Development and Deployment of AI for Mental Health and Life cycle Habits Recommendation System

A Capstone Project report submitted
in partial fulfillment of requirement for the award of degree

BACHELOR OF TECHNOLOGY
in
SCHOOL OF COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE
by

Chokkam Chandu	**(2203A52081)**
Burra Gopikrishna	**(2203A52076)**
Vamshikrishna Gattu	**(2203A52127)**
Tirumalapudi Navya Lohita	**(2203A52057)**
Yathamshetty Rithwik	**(2203A52189)**

Under the guidance of
Dr Amit Kumar Yadav
Assistant Professor, School of CS&AI.

SR University, Ananthsagar,Warangal,Telagnana-506371

CERTIFICATE

This is to certify that this project entitled **“Development and Deployment of AI for Mental Health and Life cycle Habits Recommendation System** " is the bonafide work carried out by**, Chokkam Chandu, Burra Gopikrishna, Vamshikrishna Gattu, Tirumalapudi Navya Lohita, Yathamshetty Rithwik** as a Major Project for the partial fulfillment to award the degree **BACHELOR OF TECHNOLOGY** in **School of Computer Science and Artificial Intelligence** during the academic year 2025 2026 under our guidance and Supervision.

Dr Amit Kumar Yadav	**Dr. M. Sheshikala**
Assistant Professor	Professor & Head,
SR University	School of CS&AI,
Anathasagar, Warangal	SR University
	Ananthasagar, Warangal.

Reviewer 1	**Reviewer 2**
Name:	Name:
Designation:	Designation:
Signature:	Signature:

ACKNOWLEDGEMENT

We owe an enormous debt of gratitude to our Major project guide **Dr Amit Kumar Yadav, Assistant Prof** as well as Head of the Department of CS&AI, **Dr. M. Sheshikala** for guiding us from the beginning through the end of the Major Project with their intellectual advice and insightful suggestions. We truly value their consistent feedback on our progress, which was always constructive and encouraging and ultimately drove us in the right direction.

We express our thanks to project coordinators **Mr. Sallauddin Md, Asst. Prof.** for their encouragement and support.

Finally, we express our thanks to all the teaching and non-teaching staff of the department for their suggestions and timely support.

ABOUT THE ORGANIZATION

SR University (SRU), an autonomous private institution located in Warangal, Telangana, India, offers a variety of academic programs with an emphasis on interdisciplinary learning and practical, real-world application. The university provides both undergraduate and postgraduate courses, including B.E/B.Tech, MBA/PGDM, M.E/M.Tech, BBA, B.Sc, and M.Sc programs. SRU's curriculum encourages innovative thinking and the integration of theoretical concepts into practical solutions.

The university boasts excellent infrastructure, featuring a large library and a digital library, a sports complex, and modern hostels for both male and female students. SRU also offers flexible academic options such as a flexible credit system,

opportunities for minor degrees and specializations, branch changes, integrated programs, and semester abroad plan. With a strong focus on research, SRU has over 50 sponsored projects and more than 190 patents to its name.

The university's placement cell provides robust support for job placements and internship opportunities.

SRU also fosters extra-curricular engagement through clubs like dance and music clubs, which enable students to participate in various intra and inter college competitions. Additionally,

SRU's Ph.D. program is designed to provide comprehensive training in research methodology, technical communication, and literature review.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Abstract

The importance of mental health and wellbeing as a whole has increased in recent times. This change is due to rapid changes in social behavior and work environment and daily life. People's mental health is suffering as a result of a number of factors. These include an increase in academic work, busy work schedules, insufficient sleep, excessive screen time, poor eating habits, and insufficient physical activity. Traditional techniques used for tracking mental health often are random, slow, and fail to offer personalized insights. In such a case, there is a surging need for smart and data-driven systems. These systems can help people learn more about their mental health and how they can work to improve their habits. With this motivation, the present project titled *Development and Deployment of AI for Mental Health and Life Cycle Habits Recommendation System* aims to predict a person's happiness score—ranging from 1 to 10—based on key behavioural and lifestyle attributes.

The system makes use of a structured data with demographic, psychological, and lifestyle-related variables, like age, gender, stress level, sleep duration, work hours, diet quality, physical activity, medication use, smoking and alcohol use, involvement in social media. Following required data pre-processing, a variety of machine learning regression models were trained: Linear Regression, Ridge Regression, Lasso Regression, Decision Tree Regressor, Random Forest Regressor, and XGBoost Regressor, which are considered the most appropriate of them to predict the happiness score. These models were assessed with respect to the right performance measures such as R2 score, RMSE and cross-validation. The continuous prediction of happiness offers the user a better idea of his or her mental wellbeing, and also shows how their lifestyle choices every day affect their emotional wellbeing.

Following the transparency principle and increasing user trust, the system integrates Explainable Artificial Intelligence (XAI) with LIME. This will allow users to see the impact of individual features on their happiness score that they are predicted to get, thus they can learn what habits they need to work on. Moreover, the software will have a recommendation engine in which data about the user is matched with the high-happiness trends to make customised recommendations that should enhance mental health.

The entire model is implemented on an interactive Gradio interface making it accessible to a general user. The interface makes real-time predictions and explanations and intuitive visual dashboard. All in all, the project reveals how artificial intelligence, integrated with the analysis of behaviours, can be an efficient tool of mental health support, positive lifestyle habits, and the influence on individual decision making to benefit wellbeing long-term.

CHAPTER 1
INTRODUCTION

1.1 Overview of Mental Health and Lifestyle Wellbeing

The twenty-first century has brought about the importance of mental health and lifestyle wellbeing in the life of human beings. As the academic settings, competitive work ethics, urbanisation and continuous exposure to technological appliances continue to intensify, physical and emotional health of people of all ages is changing drastically. The problems of higher levels of stress, insufficient sleep, unhealthy eating habits, reduced social contacts, and reduced physical exercise have become general over the recent years, particularly among young adults and professionals at work.

Such lifestyle habits directly impact the wellbeing of their minds, which frequently results in mood disorders, anxiety, burnout, and deterioration of general life satisfaction.
Mental health awareness and discussion in India is slowly being brought to light, yet there is a strong reluctance to openly accept the issue of mental health. A large number of people put off help or even do not seek help because of the stigma associated with the society, financial cost or because the mental health services are not readily available. In those cases, it is hard to know what is happening to ones mental health leading to poor behavioural loops which go on as they go undiagnosed.

In the meantime, artificial intelligence (AI) and machine learning (ML) have also changed a multitude of industries, such as health care, finance, manufacturing, and education. These technologies have the capability of analysing massive data, recognising trends, learning behavioural contributions, and make quality predictions. Mental health monitoring AI-based offers novel opportunities to know more about behavioural patterns and predict the outcomes of wellbeing. The happiness score is one of such measurable results; it provides an indication of the level of satisfaction or emotional balance of a person. This score has a scale of 1 to 10, which is readily understandable as an indicator of mental wellbeing.

Given the progress of AI, access to online behavioural data, and the necessity to provide personalised insights on wellbeing and lifestyle habits, the creation of an AI-based mental health and lifestyle habit recommendation system will be of high value. This project will use machine learning to predict happiness levels and suggest changes to the lifestyle that can help to achieve improved wellbeing and better mental health.

1.2 Problem Statement

The issue of mental health has increased tremendously in all corners of the society, though most people are in the dark of the role of their daily practices in affecting their mental life. In most instances, individuals do not understand that emotional balance is eroded with time as a result of sleeping irregularly, being in a state of constant stress, engaging in no physical activities, following an unhealthy diet, experiencing excessive dependence on social media, smoking, drinking, and working long hours. Because these habits become the other norm of life, people usually do not consider the long-term consequences these habits have on happiness, motivation, and clear mind. This has led to a good number of individuals living with dysfunctional mental conditions without a clue of the underlying causes.

The conventional mental health evaluation tools are very dependent on clinical assessment, counselling and psychological testing. These are effective, but they are not viable to all since not everyone can afford them. In India, particularly, the number of professionals in mental health is poor in comparison to the population. Social stigma causes many people to be reluctant to seek professional assistance because they are unaware, or they cannot afford to do so. This poses a big gap because individuals fail to get feedback or advice in time in matters concerning their wellbeing.

The other big problem is the subjectivity involved in self-assessments. Most people fail to make a correct assessment of their mindset depending on the immediate feelings or suppositions resulting in prejudiced conceptions. In the absence of objective instruments, people will be unable to properly quantify such variables as the level of happiness, the degree of emotional satisfaction, and the psychological balance. In addition, the majority of the population lacks the technological or psychological capabilities to analyse data and determine the lifestyle patterns that are harmful independently.

Although digital wellbeing apps are available, there are no intelligent systems that can use machine learning predictions, explainable AI, and personalised recommendations. Individuals rarely get elaborate descriptions on the habits that are influencing their wellbeing or how changing some behaviours can enhance their happiness. This is a shortcoming of personalisation that restricts the utility of current wellbeing tools.

Thus, it is of an urgent need to introduce a scientifically shaped, AI-powered, user-friendly system, which could anticipate happiness levels, interpret behaviour trends, draw attention to the most impactful factors, and suggest lifestyle-related changes. This kind of a system can enable the citizens to have an objective view of their mental state and make knowledgeable choices that would improve their living situations. This is a gap that will be addressed by this project, which seeks to come up with a mental health and lifestyle habit recommendation system.

1.3 Objectives of the Study

The objective of this study is to design, develop, and deploy an AI-based system that predicts happiness scores and provides personalised lifestyle recommendations to enhance mental wellbeing. The system aims to act as a supportive digital tool, offering users insights into their behavioural patterns and helping them identify areas of improvement. To achieve this purpose, the following objectives have been framed:

1. **To build a machine learning–based predictive model for happiness scoring:**
   - The primary objective is to develop a model that analyses lifestyle and behavioural patterns to predict a happiness score on a scale of 1 to 10.
   - The system uses features such as stress level, sleep duration, physical activity, diet quality, work hours, and social media habits, among others, to estimate the user's psychological state.

2. **To evaluate and compare multiple machine learning algorithms:**
   - To ensure accuracy, the study compares various regression models such as Linear, Ridge, Lasso, Decision Tree, Random Forest, and XGBoost.
   - Each model is evaluated using performance metrics like R^2 score, RMSE, and cross-validation. The aim is to identify the most stable and reliable predictor of happiness.

3. **To integrate Explainable Artificial Intelligence (LIME):**
   - Modern AI systems often work like "black boxes," making users hesitant to trust predictions. By incorporating LIME (Local Interpretable Model-Agnostic Explanations), this project ensures transparency.
   - Users can see how each lifestyle factor contributed to the predicted happiness score, helping them understand every recommendation deeply.

4. **To create a personalised habit recommendation module:**
   - Using clustering techniques, the system identifies behaviour patterns from individuals with higher happiness score and compares them with the user's profile.
   - This allows the system to offer personalised recommendations such as improving sleep routines, reducing screen time, managing stress, or increasing physical activity.

5. **To deploy the complete solution through a user-friendly interface:**
   - The Gradio interface is used to make the system easy to access. Users can enter their details, receive predictions instantly, view visual explanations, and explore lifestyle suggestions.

6. **To promote wellbeing awareness:**
    - Another important objective is to help individuals become more conscious of their mental health and encourage them to adopt healthier behaviours.

Together, these objectives aim to create a comprehensive, accessible, and intelligent wellbeing support system.

1.4 Scope of the Study

The scale of the research is a wide array of activities that are associated with the creation and implementation of an AI-based mental health and lifestyle recommendation platform. The project will start with the knowledge on the connection between lifestyle variables and mental health and move forward to constructing a smart system capable of processing these variables and providing useful predictions.The research involves the identification and the use of a systematic dataset that contains a set of variables according to the age, gender, level of stress, sleep, work hours, nutrition, physical activity, social media use, medication history, and substance use. It goes through pre-processing measures such as processing missing values, cleaning the data, encoding nominal variables and scaling the numerical variables to work with machine learning modelling.Other areas that come in the scope include application of different regression models of supervised machine learning to determine the most fitting model to predict the level of happiness.

These are Linear Regression, Ridge Regression, Lasso Regression, Decision Tree Regressor, Random Forest Regressor and XGBoost Regressor. To achieve accurate and valid predictions of their models, the performance of the models is measured on standard measures like R 2 score and R MSE.One of the significant aspects of the project is the implementation of Explainable AI (XAI) by the usage of LIME. This is to make it transparent so that one can see the role of one feature in the model output. Users can also visualise the negative impacts of such behaviours as lack of sleep, stress, or unhealthy foods on happiness.

Besides that, the research involves coming up with a personalised lifestyle recommendation module based on clustering methods. This component helps determine the behavioural patterns that are linked with an increase in happiness of the user and gives individualised advice to the user, which could be to eat better, spend less time on the screen, work out more or to deal with stress in a constructive way.

Lastly, the project involves the implementation of the entire system with Gradio to enable the use of the whole system by people with no technical expertise. The application provides a user-friendly interface through which users can feed data, see forecasts, get an explanation and a suggestion in real time.The research does not require the diagnosis of mental health conditions or the substitution of clinical interventions. Rather, it specializes in creating awareness, encouraging healthy habits and giving users data-informed insights in order to enhance their wellbeing.

CHAPTER 2
LITERATURE REVIEW

2.1 Studies on Happiness, Mental Wellbeing, and Lifestyle Factors

Research on happiness, mental wellbeing, and lifestyle determinants has expanded rapidly over the past few decades across multiple disciplines, including psychology, behavioural science, public health, sociology, and economics. Scholars increasingly recognise that subjective wellbeing is not merely an abstract emotional state but a comprehensive construct shaped by cognitive evaluations of life satisfaction, emotional stability, physiological health, and the fulfilment of personal and social needs. As a result, wellbeing has become an important indicator of societal progress, often considered alongside traditional economic measures such as GDP.

A significant body of literature underscores that happiness arises from a complex interplay between internal psychological characteristics and external environmental circumstances. Studies from positive psychology, beginning with the works of Martin Seligman and Ed Diener, conceptualise wellbeing as comprising positive affect, the absence of negative affect, life satisfaction, meaning, and personal growth. Similarly, behavioural scientists emphasise that routine lifestyle choices—including sleep habits, nutrition, physical activity, substance use, and daily stress exposure—directly influence emotional resilience and long-term happiness outcomes.One of the most influential and comprehensive resources in this domain is the *World Happiness Report*, published annually and widely used to assess global wellbeing levels.

The report consistently identifies six major predictors of national happiness levels:

- **Social support and interpersonal trust**
- **Economic security and stable income**
- **Healthy life expectancy and access to healthcare**
- **Freedom to make personal life decisions**
- **Generosity and prosocial behaviour**
- **Trust in institutions and societal integrity**

These indicators reveal that wellbeing is deeply multidimensional, encompassing social, economic, psychological, and environmental factors. The global patterns reported each year show that individuals who have strong social connections, stable health, and autonomy experience significantly higher levels of life satisfaction. The report also stresses that societal context influences personal happiness as strongly as individual behaviour does.

At an individual level, lifestyle choices have been extensively studied for their effects on emotional health. Numerous empirical studies highlight that insufficient sleep, chronic stress, physical inactivity, irregular work schedules, excessive screen exposure, and unhealthy eating habits have detrimental effects on mood regulation and overall mental wellbeing. Mental health research demonstrates that disrupted circadian rhythms and sleep deprivation significantly heighten the risk of anxiety, irritability, cognitive impairment, and reduced happiness levels. Similarly, constant exposure to digital devices and social media is associated with emotional fatigue, increased self-comparison, reduced attention spans, and poorer life satisfaction—especially among adolescents and young adults.

On the other hand, adaptive lifestyle behaviours serve as protective factors that enhance mental stability. Studies in behavioural medicine consistently show that regular physical activity improves mood, reduces cortisol levels, enhances neurochemical balance, and contributes to sustained happiness. Nutritional research further indicates that balanced diets rich in fruits, vegetables, proteins, and micronutrients support emotional regulation, energy levels, and mental clarity. Healthy sleep patterns restore physiological equilibrium, improve cognitive functioning, and enable emotional resilience. Social psychology research also affirms the critical role of interpersonal relationships, emotional support networks, and community belonging in safeguarding wellbeing.

Many scholars categorize these lifestyle determinants of wellbeing into the following broad clusters:

- **Physical health variables:** sleep quality, exercise frequency, diet, hydration
- **Psychological variables:** stress levels, emotional regulation, mindfulness habits
- **Social variables:** relationships, social bonding, community support
- **Work-related variables:** work hours, job satisfaction, work-life balance, burnout
- **Behavioural risk factors:** smoking, alcohol consumption, digital addiction

Emerging evidence also highlights that contemporary digital-age stressors—including increased online engagement, reduced physical activity, and high cognitive load—are significantly altering happiness trajectories among working professionals, students, and urban populations. In India, for instance, rapid societal transitions, competitive academic and employment environments, long working hours, and urban lifestyle pressures have contributed to rising stress levels and declining wellbeing metrics. Surveys by national psychological associations indicate a rise in burnout, emotional exhaustion, and dissatisfaction with life among young adults and middle-aged individuals, underscoring an urgent need for accessible wellbeing assessment tools.

While the body of research on happiness and mental wellbeing is extensive, a notable gap persists in the literature: **the majority of studies focus on correlation and statistical interpretation rather than providing personalised, actionable guidance for individuals.** Similarly, although modern AI and machine learning technologies have been applied widely in healthcare prediction and behavioural analytics, their use in personalised wellbeing prediction and tailored lifestyle recommendations remains relatively limited.
This gap highlights the necessity for intelligent systems capable of analysing individual behavioural patterns, identifying lifestyle risks, predicting wellbeing outcomes, and delivering adaptive, personalised recommendations. Such systems bridge the divide

between theoretical research and practical intervention—precisely the objective undertaken by the present project.

2.2 Machine Learning Approaches in Wellbeing and Happiness Prediction

Machine Learning (ML) has emerged as one of the most transformative analytical paradigms in contemporary wellbeing research. Unlike traditional statistical approaches that rely heavily on predefined assumptions, linearity constraints, and rigid model structures, ML methods possess the capability to learn highly non-linear, multivariate, and context-dependent patterns directly from data. This flexibility makes them exceptionally suitable for modelling complex human behaviours, psychological responses, and lifestyle-driven fluctuations in subjective wellbeing. Over the past decade, ML has been increasingly utilised to analyse behavioural data, predict affective states, diagnose early indicators of mental health disorders, and estimate subjective happiness with growing precision.

A significant milestone in this domain is the work of Oparina et al. (2025), who conducted a large-scale investigation across Germany, the United Kingdom, and the United States using extensive wellbeing datasets. Their empirical findings revealed that ML algorithms such as Random Forests, Gradient Boosting Machines, and LASSO Regression markedly outperformed traditional Ordinary Least Squares (OLS) models in predicting life satisfaction. The superiority of ML arises primarily from its ability to capture non-linear associations, threshold effects, and interaction patterns among influential variables such as social support, income stability, health quality, and network size. Although the reported upper-bound correlation ($R \approx 0.30$) reflects the inherent difficulty of modelling subjective wellbeing, it also underscores the substantial explanatory value offered by machine learning over classical methods.

Similarly, Celik et al. (2025) extended this line of research by applying ML techniques to the World Happiness Index dataset. Their comparative evaluation included Logistic Regression, Decision Trees, Artificial Neural Networks (ANN), Support Vector Machines (SVM), and the XGBoost algorithm. The study demonstrated that although linear models and SVMs performed competitively, the performance of XGBoost was comparatively lower in this specific dataset. This reinforces an important insight in ML research: model performance is heavily dependent on data characteristics, including feature distribution, noise levels, sampling variability, and the dimensional complexity of the underlying patterns. No single algorithm performs best across all datasets, highlighting the necessity of systematic experimentation and model selection.

Beyond large population-level analyses, machine learning has been increasingly applied to domain-specific mental health and wellbeing assessments. For example, ML models have been employed to:

- detect stress and behavioural anomalies using smartphone sensor data such as screen time, app usage, keystroke dynamics, and mobility patterns;
- identify early symptoms of depression through lexical and semantic analysis of social media posts;
- forecast emotional states using physiological data collected from wearable sensors, including heart rate variability, sleep cycles, and galvanic skin response;
- classify happiness, anxiety, or emotional arousal using voice tone, speech rhythm, and facial expression datasets.

These applications highlight the versatility and adaptability of machine learning in processing heterogeneous behavioural data and deriving psychologically meaningful insights.

However, despite these advancements, ML-based wellbeing prediction encounters several challenges. Foremost among them is the issue of interpretability. Many of the most accurate ML models—such as Deep Neural Networks, Gradient Boosting frameworks, and ensemble architectures—operate as black-box systems, offering limited transparency into how specific input features influence the final output. This opacity poses significant ethical and practical concerns, particularly in sensitive fields like mental health, where users require clarity and justification for algorithmic outcomes. Lack of interpretability reduces user trust, limits adoption, and can even lead to misinterpretation of model predictions.

To address these challenges, researchers have increasingly turned toward Explainable Artificial Intelligence (XAI) methodologies, most prominently LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (SHapley Additive explanations). These techniques enable meaningful insight into the relative contribution of individual features—such as stress levels, sleep duration, diet quality, or work hours—to predicted happiness scores. Such interpretability bridges the gap between predictive accuracy and user comprehension, making ML models more aligned with ethical, transparent, and human-centric AI principles.

In summary, machine learning has made substantial contributions to the prediction of subjective happiness and psychological wellbeing across various populations and data modalities. Yet, much of the current academic work remains focused solely on prediction rather than on explainability, personalisation, and actionable guidance. The present project aligns with the emerging need to integrate prediction, interpretability, and tailored recommendations into a unified platform. By combining ML modelling with XAI techniques and a personalised recommendation engine, the system developed in this project advances the field toward a more meaningful, user-centered, and practically deployable approach to wellbeing analysis.

2.3 Limitations of Existing Wellbeing Assessment and Digital Mental Health Systems

Despite the availability of various tools for monitoring lifestyle behaviours and mental health, most existing systems suffer from substantial limitations that reduce their real-world usability and overall effectiveness. Traditional wellbeing assessment methods continue to rely heavily on manual surveys, psychological scales, periodic check-ups, and clinical interviews. Although these techniques are validated and widely used, they exhibit several critical drawbacks that hinder timely and personalised wellbeing management.

Limitations of Traditional Wellbeing Assessment:

- **Time-consuming procedures:** Standard clinical assessments require scheduled appointments, long questionnaires, and professional interpretation, making them impractical for continuous monitoring.

- **Subjective and biased responses:** Self-reported survey answers are vulnerable to mood biases, memory distortions, and social desirability effects, reducing result reliability.

- **Lack of ongoing, real-time evaluation:** Traditional tools provide only occasional snapshots of wellbeing, failing to capture day-to-day emotional fluctuations or behavioural trends.

- **Limited accessibility:** Regular assessments can be expensive, geographically restricted, and require psychological expertise that many users do not have access to.

- **Stigma and reluctance to seek help:** Many individuals avoid clinical or therapeutic environments due to fear of judgement, leading to underreporting of mental stress.

While digital wellbeing and mental-health apps attempt to address these issues, they also present significant limitations. Most applications focus on tracking isolated lifestyle metrics—such as step count, sleep duration, or calorie intake—without integrating them into a holistic mental-wellbeing model. As a result, they lack analytical depth and fail to draw meaningful insights from multi-dimensional behavioural data.

Key Drawbacks of Current Digital Wellbeing Systems

1. Lack of Personalisation

- Many apps deliver generic, one-size-fits-all advice.
- Recommendations are not derived from individual behavioural patterns or long-term lifestyle trends.
- Users often receive repetitive or irrelevant suggestions, reducing engagement and perceived usefulness.

2. Absence of Happiness or Emotional-Wellbeing Prediction

- Very few tools quantify happiness or subjective wellbeing as a measurable score.
- Users are unable to track emotional health changes or understand what factors influence their happiness levels.
- The lack of predictive capability limits long-term wellbeing planning.

3. Lack of Explainability in Predictions

- Even when apps incorporate predictive analytics, they rarely explain how a result was generated.
- Black-box outputs reduce user trust, especially in sensitive domains like mental health.
- Without transparency, individuals cannot identify which behaviours require modification.

4. Fragmented and Narrow Data Tracking

- Many platforms track only one or two parameters (e.g., sleep-only, steps-only, or calorie-only apps).
- Mental wellbeing is multi-factorial, influenced by stress, relationships, habits, work hours, and lifestyle routines—factors rarely captured together.
- This fragmented data makes holistic wellbeing analysis impossible.

5. Limited or No Machine Learning Integration

- Most digital wellbeing systems rely on rule-based heuristics rather than data-driven ML models.
- They cannot detect hidden patterns, subtle correlations, or complex non-linear behaviours.
- As a result, the insight quality is low and fails to reflect the dynamic nature of human wellbeing.

6. Absence of a Recommendation Engine

- Current solutions often provide static, non-adaptive advice.
- They do not compare user behaviour against patterns associated with happier individuals.
- No behavioural clustering or personalised recommendation logic is applied.

7. Poor Cultural Adaptation and Limited Context Relevance

- Most tools are built using Western datasets and psychological assumptions.
- Cultural differences in habits, stressors, family structure, and lifestyle behaviours significantly affect wellbeing but are often ignored.
- For users in regions like India, many suggestions feel irrelevant or misaligned with local contexts.

These limitations highlight the urgent need for a system that achieves:

- Data-driven happiness prediction using machine learning
- Explainable outputs using XAI techniques like LIME or SHAP
- Personalised and adaptive lifestyle recommendations
- A culturally relevant and user-friendly design
- Holistic tracking of multiple behavioural features

The present project addresses these gaps by integrating ML-based prediction, interpretability, and personalised behavioural recommendations into one unified and accessible wellbeing platform.

CHAPTER 3
METHODOLOGY

3.1 Dataset Description

A closely related issue is the need to have strong and deep data of various kinds with good statistical reliability to train and test the data-driven machine learning system. In a wellbeing prediction platform -- predicting the self-reported happiness score of a user based on lifestyle patterns, psychological states, demographic information, etc. -- the significance of dataset becomes apparent. Because Wellbeing is a highly subjective, multi-dimensional and non-linear construct, the data's features should be rich enough to allow capturing such subtle nonlinear interactions.

The data are a collection of variables that describe demographic, behavioral, and other proclivities related to lifestyle, physiological, and mental health. Combined, these give us a holistic view of human daily life. Such a dataset is crucial not only to achieve high predictive accuracy but also to produce interpretable and actionable results so that the derived interpretation and recommendations can in turn induce a change in real human behaviour.

3.1.1 Dataset Description and Size

The data set possesses multiway structure with the categorical features being considered as the modes and the numeric features as the slabs. This duality is suitable, as we can express human life quantity in terms such age or amount of sleep, but also life has qualitative elements such as if one has certain mental health conditions, is a smoker, or has an occupation. A complete preprocessing is required for mixed-type features, while it enriches the data set and make the model to learn on diversified behaviours.

The dataset's representativeness is another strong point. They represent a range of ages and work and life styles. The diversity in populations used in this guarantee that the training model did not overfit to a particular user group, which is further verified by better generalisation to real user inputs.

3.1.2 Demographic Features

Demographic layer the demographic layer is comprised of general information about the participants. Below are some of them:

Age:

Age plays a role in influencing lifestyle, e.g., sleep duration, work schedule, physical activity, and mental health stabilities. They find undergraduates "sleep under a rock," thinking about roadblocks in the job market and putting in long hours; those farther along in life are more established. Due to the rich age representation in the dataset, the happiness model profitts from capturing the happiness variance over life stages.

Gender:
Working hours, social media and diet trends are some of the behavioural trends which show a dependency on gender. For instance, some studies have shown differences in the way we cope, our quality of sleep and digital behaviors by gender. To account for this variation, the authors introduce the random variable z, representing the happiness state of the society at large.

Occupation:
The life of an employee is influenced in some degree by the working hours and the stress level of his daily work. Stressful occupations (IT, finance, health care) might be associated with unhealthy practices such as less sleep and more stress, whereas occupations with more flexibility or creativity might be associated with more positive behaviors. Country/Region: Differences among regions in traditional culture, work ethic, eating habit, social culture, and digital usage are very obvious. The model can also utilize geographical information to enable better generalization to different societal settings. Together, these demographic variables act as a baseline context to both behavioral and psychological functioning.

3.1.3 Lifestyle Traits and Behaviours

Everyday activities and behaviours (lifestyle variables) are the unique types of behaviours that directly influence day-to-day functioning and mental health are the constituent behaviours of lifestyle. Key features include:

Sleep Hours:
Not many other aspects of lifestyle/metabolism that have been so strongly scientifically linked to mood, cognitive function and well-being. Chronic sleep deprivation is associated with increased stress, anxiety, depression and decreased work productivity. This gives the model a more robust measurement of the sleep wellness relation.

Work Hours:
High demands and long working hours are risk factors for burnout, emotional exhaustion and job dissatisfaction. Excessive preoccupation with work is thought to be one of the most significant risk factors (Albrecht 2012) the time pressure resulting from highly demanding work schedules may reduce the time available for sleep, socialising and other activities that have beneficial effects on life satisfaction.

Physical Activity Hours:
Physical activity is a consistent predictor of mood enhancement-results from activity on mood are partly attributed to physiological factors such as release of endorphins, increase in quality of sleep, and decrease of stress. It also provides additional detail on the salutogenic effects of fitness on wellbeing.

Social Media Usage:
The complex effects of social media on mental health in the digital generation are among the primary barriers to expanding diagnosis and care capacity globally. Social media has a moderate benefit for social connection, but intense use has been linked to anxiety, stress from social comparison, and decreased happiness. The range depicted is to show the extremes of these behaviours.
Given these two variables, the model player can determine clusters of positive and negative life styles; and use that to evaluate the effect on well-being.

3.1.4 Psychological and Clinical Characteristics

Cognitive characteristics complicate the modeling since attended happiness is not just sensed externally—it is embedded in one's mind. Among the relevant variables are the following:

Mental Health Condition:
Indicates whether the user has any diagnosed mental illness (similar anxiety, depression). This changes everything in terms of how they respond to stress, how they sleep, how they work, how they live their lives.

Severity:
Indicates the degree of the mental disorder. Severity levels enable the model to separate mild stress-related conditions from extreme disorders.

Consultation History:
The person has never sought professional help for mental illness. This speak to some level of knowledge, coping and treatment related encounter.
Clinical factors should be included as patterns of behaviour may be markedly different in those with latent psychiatric disorder.

3.1.5 Additional Behavioural Indicators

The dataset also has a number of health-related behavioural attributes:

Stress Level – possibly the most potent and negative predictor of happiness.
Diet Quality– has an impact on the physical health, energy, and emotional stability.
Smoking Habit & Alcohol Consumption– are coping strategies in the short term, but risk to well-being in the long term.
Drug Use:
signifies an chronic condition or mental/physical health problem.
Together, these variables constitute a dynamic behavioural representation that reflects the complexity of humans in the real world."

3.2 Data Pre-processing and Feature Engineering

Data processing and feature engineering is critical in any machine learning pipeline, and is particularly crucial for human behaviour psychology, life monitoring and mental health analytics applications. Raw data sets about the human are susceptible to noise, subjective bias, missing data, and heterogeneous data formats, as well as a wide array of behaviours. Without proper handling, these issues can severely skew data distributions, mislead the learning algorithm to yield less accurate, less stable and less fair prediction models. Therefore, a multi-stage pre-processing pipeline was developed for this study to transform raw lifestyle data into a clean, reliable, and analyzable format.

The primary goal of this pre-processing stage pipeline is to achieve (i) data completeness, consistence, interpretability, and model friendliness; (ii) behavioral diversity, and natural human habit. Here, we explain in detail the cascade of transformations used, the rationale for each step, and the contributions of these procedures to the final robust happiness prediction model.

3.2.1 PROCESSING OF MISSING VALUES

Missing data is one of the most frequent issues for well-being-data related datasets. Due to privacy and embarrassment, participants may decline to answer sensitive questions (mental health history, alcohol consumption, level of stress, use of medication). Additional missing entries may be brought on by survey fatigue or incomplete digital logs completion). These gaps present a major problem – one cannot train machine learning models directly on missing data, and unmitigated coverage gaps may also bias or destabilize training.

A column-wise analysis was performed2 to assess the extent of missingness. Numerical variables exhibited moderate missingness, (e.g. Sleep Hours, Work Hours, Stress Level Physical Activity Hour), but categorical variables for missing numerical observations, mean or median imputation was employed based on the skewness of the distribution. Median imputation was used instead for heavily skewed social variables (ex, since mean is a non-robust estimator and it is affected by outliers) as mean was not used for these variables.

Mode-based imputation (most frequent class) was performed for categorical variables. Although this does not generate novel behavioral patterns, it does maintain category integrity and helps to reduce sparsity. Other imputation methods (for example KNN imputation) were considered but not selected for the final pipeline since they would artificially increase similarity among participants, which may not be wanted in behavioural-modelling.

Consequently, a processing was conducted to ensure that rows were not removed due to missing features, meaning that no valuable data samples were discarded.

3.2.2 ENCODING OF CATEGORICAL VALUES

Machine learning models accept only numeric values as input. Therefore, categorical features were transformed to numeric values by applying suitable encoding schemes.

a**) Ordinal Encoding**

There are some life style and behaviour indicators which are naturally ordered.
Example:

Diet Quality (Healthy > Moderate > Poor)
Severity (None < Mild < Moderate < Severe)

Since the order of the categories represented an important information, Label Encoding were used to encode those features. By maintaining this ordinal nature, models can be trained to predict progressive behavioural decline or improvement.

b) One-Hot Encoding

Categorical variables without any intrinsic ordering were one hot encoded and converted into several binary variables. This was applied to:

- MentalHealthCondition
- SmokingHabit
- AlcoholConsumption
- Occupation
- Country/Region

One hot encoding doesn't imply a fake hierarchical order such as one category being "greater" than another.This sharp distinction of ordinal and nominal features helped to keep the semantic meaning of each feature as a number.

3.2.3 OUTLIER DETECTION AND CORRECTION

Outliers are those values in numerical data where the observations are extreme, meaning the data value in those observations deviates so much from normal behaviour. Diet and activity data may have inflated or otherwise unrealistic records, such as:

- Work Hours is over 20 per day
- Sleep Hours is close to 0
- Social Media Usage is more than 15 hours.

Furthermore, additional EDA (eg, boxplot, distribution plot, Z-score) was carried out to identify these outliers. They can also be caused by data entry errors, overzealous or extreme self-reporting, or possibly just a very rare but valid human act.

So, to align them:

- A few thresholds were added (e.g., max biologically plausible sleep hours = 14) informed by domain knowledge.
- The unrealistic values were replaced or removed.

- Genuine extremes (e.g., high stress and low sleep) were retained as they add richness to the dataset.

This trade off guaranteed that noise- based distortions were removed at the cost of minimal loss of genuine behavioural features.

3.2.4 Scaling and Normalization of Data

Numerical features of the data are on very different scales:

- Work Hours ranges from 0 to 16
- Social Media Usage ranges between 0 and 10
- Stress Level is from 1 to 10
- Sleep Hours: The value is between 2 and 14

Features with high numeric values may dominate model training when unscaled, particularly distance based or gradient based models (e.g. KNN, SVM, LR).
Two approaches to scaling were investigated:

a) Standard Scaling:

For those features that are roughly normally distributed:

- Stress Level
- Sleep Hours
- Physical Activity Hours

It scales the data to have zero mean and unit variance.

b) MinMax Scaling :

Applies to skewed features or features that are naturally bounded: Work Hours, Social Media Usage , Diet Quality (after encoding).

This alters all the values to fit into the range [0, 1], this was done so that they would contribute evenly. The normalization guaranteed that no one exercise would overwhelm the prediction engine.

3.2.5 Feature engineering

We applied feature engineering to derive more behavioral features.

a) Sleep–Stress Interaction Score

A corollary feature combining:

- Sleep Hours
- Stress Level

This reflects a sleep-stress disparity which is a very predictive measure of happiness. Stress and sleep deprivation combined are a toxic brew associated with ill health.

b) Lifestyles Balance Score

Add up:

- Physical Activity Hours
- Diet Quality

Lifestyle Composite Score The composite score represents a more holistic view of lifestyle well-being. Individuals who consume a healthy diet and regularly participate in physical activity have enhanced mood and higher levels of reported happiness.

3.2.6 Dataset Splitting

After the cleaning and transformation processing, the clean datasets were split into the following groups:

- 80% Training Data
- test data, 20% of the data overall

This partitioning ratio enabled for the following:

- That's train enough
- We have enough unseen data to get a sustainable evaluation.

The test set provides a second test to obtain an estimate of the generalisation performance.

3.2.7 Pipeline integration Pre-processing

All the operations required to transform the data (imputation, encoding, scaling, and feature engineering) were performed using a unified pre-processing pipeline. That is to guarantee

- train and user-input data are always transformed the same way
- To be reproducible when deployed
- There is no data leakage (fit the transformation on train data only)

The joint pipeline is critical for production because users give input values that must be processed in the same way before they are passed to the model.

3.3 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) forms one of the most essential and intellectually foundational phases in any data-driven research methodology, particularly when the domain involves human lifestyle patterns, behavioural dynamics, and mentally sensitive wellbeing indicators. In the context of this project—which aims to predict happiness scores through a comprehensive analysis of lifestyle attributes—EDA acts as the cornerstone for understanding latent relationships, distribution structures, behavioural anomalies, and underlying statistical regularities embedded within the dataset. Unlike purely technical preprocessing steps, EDA engages both quantitative assessments and conceptual reasoning to produce a holistic understanding of how variables interact, influence, or distort one another in shaping the overall outcome variable, i.e., the happiness score.

Given the rich multi-dimensionality of the dataset, comprising numerical, categorical, behavioural, and psychological variables, EDA was executed through an interconnected sequence of statistical summaries, visual techniques, multivariate pattern analyses, and diagnostic explorations. These analytical lenses did not merely enhance familiarity with the dataset structure but also enabled us to verify hypotheses, challenge assumptions, and ultimately inform the modelling and feature-engineering strategies used later in the pipeline.

3.3.1 Distributional and Statistical Profiling of Numerical Variables

The preliminary stage of EDA involved generating fundamental distribution plots such as histograms, kernel density estimation curves, and frequency plots for all numerical variables, including SleepHours, StressLevel, PhysicalActivityHours, WorkHours, and HappinessScore. Statistical measures—mean, median, variance, skewness, kurtosis, and interquartile ranges—were computed to capture both central tendency and dispersion.

HappinessScore Distribution

The distribution of HappinessScore displayed no pronounced skew, indicating that the population sample includes individuals from a wide spectrum of emotional wellbeing. This uniformity is beneficial for modelling because it prevents the target variable from being biased toward only low or high happiness categories.

SleepHours Distribution:

SleepHours exhibited mild right skewness, revealing that a substantial subset of individuals receives less sleep than medically recommended. From a behavioural science perspective, this is consistent with contemporary studies showing that a lack of sleep is among the strongest predictors of emotional instability and stress.

StressLevel Distribution:

The variable StressLevel demonstrated greater variability than originally expected. The presence of a long-tailed distribution suggested that while most individuals experience moderate stress levels, a smaller yet significant portion faces intense psychological distress—likely contributing to the depression of happiness scores.

WorkHours Distribution:

WorkHours displayed a multi-modal distribution, hinting that participants operate across distinct occupational or lifestyle clusters (e.g., students, full-time professionals, shift workers).

The implications of such heterogeneity required deeper subgroup analysis conducted later in the EDA.

This granular profiling offered insights into data transformations required (e.g., normalization, scaling) and revealed inherent lifestyle patterns associated with wellbeing.

3.3.2 Outlier Examination via Box Plots and Variance Profiling

Outliers represent values that deviate substantially from the general trends of the dataset. In the wellbeing context, outliers may represent either data inconsistencies or legitimate extreme behavioural states (e.g., overworking, excessive digital engagement, severe lack of sleep).

Box plots enabled clear visualization of the extremes in variables such as:

- WorkHours (extremely high values suggest burnout-prone behaviour)
- SocialMediaUsage (upper outliers indicate excessive screen dependency)
- SleepHours (lower outliers reflect chronic sleep deprivation)
- StressLevel (higher outliers reveal severe emotional or occupational strain)

A careful distinction was made between:

1. True behavioural outliers (real reflections of atypical but valid lifestyle patterns)
2. Erroneous data outliers (values likely due to entry errors, misunderstanding of questionnaires, or sensor inaccuracies)

This differentiation was essential, as eliminating true behavioural extremes would reduce ecological validity, while retaining erroneous ones would distort the machine learning model's internal representation of relationships.

3.3.3 Relationship Analyses through Scatter Plots

Scatter plots offered visual interpretability of the bivariate relationships between independent variables and the happiness score.
Key observed relationships include:

- SleepHours vs. HappinessScore: A visibly positive trend emerging from the scatter cloud indicates that better sleep quality and duration are strong positive predictors of emotional wellbeing.
- StressLevel vs. HappinessScore: A sharp negative linear pattern reveals that increases in stress uniformly lead to decreases in happiness.
- PhysicalActivityHours vs. HappinessScore: A positive upward trend further supports theories in health psychology linking regular physical activity with optimism, reduced anxiety, and better mental resilience.
- SocialMediaUsage vs. HappinessScore: Higher usage tended to correlate with lower happiness scores, aligning with empirical findings on digital fatigue and social comparison.

These relational insights provided foundational justification for choosing certain models that capture nonlinear behaviours more effectively (e.g., Random Forest, Gradient Boosting).

3.3.4 MULTIVARIATE CORRELATION MATRIX AND HEATMAP INTERPRETATION

One of the most analytically revealing tools in EDA was the correlation matrix, visualized through a heatmap to emphasize strength and direction of relationships.

Key Findings

- Strongest Negative Correlation: StressLevel showed the highest inverse correlation with HappinessScore, reinforcing stress as the single most influential detrimental factor.
- Moderate Positive Correlations: SleepHours, PhysicalActivityHours, and DietQuality all displayed positive associations with *HappinessScore*, marking them as central protective wellbeing factors.
- Negative Correlations with Digital and Work Factors: WorkHours and SocialMediaUsage demonstrated modest negative correlations, consistent with burnout-related theories and digital dependency literature.

Additionally, several inter-feature correlations emerged:

- SleepHours ↔ StressLevel: More sleep tends to reduce stress.
- SocialMediaUsage ↔ StressLevel: Heavy social media consumers reported higher stress.
- PhysicalActivityHours ↔ DietQuality: These behaviours co-occur, forming a "healthy lifestyle cluster."

These structural insights informed feature selection and warned against potential multicollinearity that could impair linear models.

3.3.5 Behavioural Grouping and Cluster-Based EDA

To understand behavioural phenomena more deeply, individuals were grouped into clusters based on:

- Stress Level categories
- Diet Quality groups
- Work Hour brackets
- Sleep-hour adequacy profiles

Key Group-Level Observations

- High-stress individuals: Low sleep, high social media usage, reduced physical activity.
- Healthy lifestyle individuals: Higher physical activity, better diet, stable sleep, and significantly higher happiness levels.
- High-workload individuals: Exhibit lower sleep, elevated stress, reduced happiness.

These cluster-driven insights later informed more sophisticated modelling strategies and recommendation logic.

3.3.6 EDA Insights Informing Feature Engineering and Model Development

The EDA revealed multiple transformation potentials, such as:

- Creating combined features like sleep–stress imbalance and activity–diet synergy.
- Identifying variables with weak explanatory power for possible elimination.
- Recognising nonlinear variable interactions, guiding the choice of tree-based models.

EDA thus played a strategic role in shaping how data was ultimately fed into machine learning algorithms.

3.4 Machine Learning Models and Training Procedure

The construction of a reliable predictive system for estimating an individual's happiness score necessitates a modelling strategy capable of navigating the intricate interplay of behavioural, demographic, and psychological determinants. Happiness, as conceptualised in contemporary wellbeing literature, is not governed by a single dominant factor but emerges from a dynamic integration of lifestyle choices, cognitive-emotional states, social behaviours, and environmental influences. Consequently, the modelling framework adopted for this study was designed to evaluate multiple machine learning algorithms, each representing distinct theoretical assumptions, structural complexities, and capacities for capturing non-linear relationships.

The overarching rationale was to avoid premature assumptions about the nature of interactions within the dataset and instead allow empirical evidence, obtained through systematic experimentation, to guide the model selection process. linear models, regularised models, hierarchical tree-based models, and ensemble boosting techniques were rigorously evaluated under a structured training and validation pipeline. Ultimately, the Extreme Gradient Boosting Regressor (XGBoost) demonstrated superior predictive performance, generalisation, and robustness, leading to its selection as the final deployment model.

The following subsections provide an expanded academic discussion of each model employed in the study, its relevance to behavioural analytics, and its comparative performance in predicting subjective happiness.

3.4.1 LINEAR REGRESSION (BASELINE MODELLING)

Linear Regression was implemented as the baseline model to establish a foundational reference point for performance comparison. It represents one of the most widely used modelling techniques in behavioural and social science research due to its interpretability and transparent structure. By assuming direct, proportional relationships between predictors such as StressLevel, SleepHours, DietQuality, or WorkHours and the happiness outcome, the model provides an initial understanding of how individual features might influence wellbeing in isolation.

However, the simplistic nature of Linear Regression presents substantial methodological shortcomings in the context of wellbeing prediction. Human behaviour rarely follows linear pathways; instead, it involves threshold effects, cumulative influences, non-linear behavioural changes, and conditional relationships. For instance, the effect of SleepHours on happiness may depend significantly on stress levels, and the positive influence of physical activity may only manifest beyond a certain duration or intensity threshold. Linear Regression lacks the capacity to model such dependencies, making it insufficient for capturing the nuanced realities of human wellbeing.

Nonetheless, Linear Regression served as a vital preliminary tool. It enabled early detection of general trends, helped identify broad directional associations, and offered a benchmark against which more complex models could be fairly evaluated.

3.4.2 Ridge Regression (L2-Regularised Linear Model)

Ridge Regression was incorporated to mitigate one of the common challenges found in behavioural data—multicollinearity. Lifestyle and psychological variables often overlap in meaning or influence, generating correlated predictors.
For example:

- Individuals with longer work hours may also report heightened stress.
- Higher social media usage may correlate with reduced sleep quality.
- Better diet patterns may coincide with increased physical activity.

Ridge Regression applies a regularisation mechanism that discourages overly large coefficients, thereby stabilising the model and reducing sensitivity to data fluctuations.

This characteristic made Ridge Regression an important extension of the baseline linear approach. It provided smoother, more generalisable predictions and reduced overfitting tendencies that can occur when multiple predictors share overlapping variance.

Despite these advantages, the model still could not account for non-linear relationships or hierarchical behavioural interactions. While it demonstrated improved stability compared to ordinary Linear Regression, its linearity constraint still limited its effectiveness in modelling complex wellbeing phenomena.

3.4.3 Lasso Regression (L1-Regularised Linear Model and Feature Selector)

Lasso Regression extends the regularisation concept by introducing a mechanism capable of performing feature selection during the training process. This property is especially valuable in wellbeing research, where not all behavioural factors may contribute meaningfully to happiness and where some variables may introduce noise or redundancy.
Lasso consistently retained core behavioural predictors such as:

- StressLevel
- SleepHours
- PhysicalActivityHours
- DietQuality

while suppressing less influential or weakly associated variables. This provided significant interpretive utility. By observing which variables were retained or reduced, the model offered insights into the behavioural indicators most strongly associated with happiness within the dataset.

However, similar to Ridge Regression, Lasso's reliance on linear associations limited its capacity to model the complex, layered dependencies inherent to subjective wellbeing. Nonetheless, its feature-selection behaviour made it a crucial component of the modelling pipeline, contributing theoretical clarity and helping refine the input space for subsequent models.

3.4.4 Decision Tree Regressor (Hierarchical Non-Linear Model)

The Decision Tree Regressor represented the first major shift away from linear modelling assumptions. Trees partition the dataset recursively, learning decision rules that form interpretable hierarchical structures.

This made the model particularly useful for capturing conditional behavioural interactions such as:

- High happiness is likely when stress is low **and** sleep is adequate.
- Happiness decreases significantly only when excessive work hours co-occur with low physical activity.
- SocialMediaUsage may be problematic primarily for individuals with already elevated stress levels.

Decision trees can naturally represent such context-dependent behaviours, aligning closely with psychological theories of wellbeing. They offer interpretability and transparency that are often lacking in more complex models.

However, their major limitations include overfitting sensitivity, especially in datasets containing behavioural outliers or non-representative lifestyle extremes. Because trees attempt to fit the training data precisely, their performance on unseen data may fluctuate. Although Decision Trees captured non-linearities effectively, they lacked the ensemble stability required for deployment-grade reliability.

3.4.5 Random Forest Regressor (Bagging-Based Ensemble Model)

Random Forests build upon Decision Trees by aggregating predictions from multiple trees trained on different subsets of data and features. This ensemble technique reduces variance, increases model stability, and significantly improves generalisation.

In the context of this study, Random Forests proved exceptionally competent in modelling:

- high-order behavioural interactions,
- non-linear relationships between predictors and wellbeing,
- variations in lifestyle patterns across individuals.

Additionally, Random Forests provided robust feature importance measures, consistently placing StressLevel, SleepHours, and PhysicalActivityHours among the strongest predictors. This aligned with existing psychological literature, reinforcing confidence in the dataset's behavioural integrity.

While Random Forests demonstrated superior performance compared to all linear models, they were computationally more intensive and still marginally less accurate than boosting-based models. Nonetheless, they played a critical role in establishing strong baseline performance for non-linear ensemble learning.

3.4.6 XGBoost Regressor (Final Model – Gradient Boosting Framework)

XGBoost emerged as the most effective model due to its ability to learn from residual errors iteratively and capture deep, layered behavioural interactions. Unlike Random Forests, which build trees independently, XGBoost builds them sequentially, with each successive tree correcting the mistakes of its predecessors. This makes the ensemble exceptionally powerful for modelling complex, interdependent lifestyle determinants.

Reasons for XGBoost's Superior Performance

1. Captures Behavioural Complexity with High Precision
Wellbeing is shaped by cumulative effects of lifestyle choices. XGBoost excels at modelling these intricate patterns due to its sequential learning mechanism.

2. Highly Effective Regularisation
Its built-in regularisation techniques significantly reduce overfitting, a major issue in behavioural datasets with variability across individuals.

3. Robust to Missing and Noisy Data
XGBoost employs efficient internal strategies for handling missing values based on learned data patterns.

4. Extensive Hyperparameter Flexibility
The ability to fine-tune depth, learning rate, subsampling, and number of trees allowed optimisation for maximum predictive accuracy.

5. Empirically Superior Across All Metrics
XGBoost outperformed all competing models on:

- R^2 Score
- MSE
- RMSE
- Cross-validation consistency

Its reliability, generalisation performance, and robustness justified its integration as the final predictive model within the Gradio interface.

3.4.6 Training and Evaluation Strategy

To ensure scientific rigor and generalisability, a comprehensive training and evaluation pipeline was implemented.

Dataset Splitting

The data was divided using an **80:20 train–test split**, ensuring the model was evaluated on unseen samples.

Evaluation Metrics

Each model was assessed using:

- R² Score – measuring variance explained
- MSE – evaluating average squared error
- RMSE – providing interpretable error magnitude

Cross-Validation

To assess stability, **k-fold cross-validation** was applied. This ensured that model performance was not dependent on a single train–test configuration.

Hyperparameter Tuning

GridSearchCV was used to tune:

- learning rate
- number of estimators
- maximum depth
- subsampling ratios
- regularisation strengths

Comparative Model Assessment

Models were compared on:

- predictive performance
- robustness
- interpretability
- consistency across folds
- computational efficiency
- suitability for deployment

XGBoost demonstrated outstanding performance across all dimensions.

3.4.8 Model Interpretability and Ethical Considerations

Given the sensitive nature of mental wellbeing prediction, interpretability and ethical transparency were considered essential.

Interpretability

- Linear and Lasso models provided coefficient insights.
- Tree-based models offered interpretable rule-based structures.
- XGBoost feature importance analyses allowed identification of the most influential behavioural factors.

Ethical Considerations

- Care was taken to avoid overgeneralisation from patterns that may not apply universally.

- The model avoids deterministic interpretations of happiness, which is inherently subjective.
- Predictions are designed to support wellbeing, not replace psychological evaluation.

3.4.9 Deployment Suitability

XGBoost was chosen for deployment due to:

- high accuracy
- robustness with real-world behavioural variation
- flexibility
- speed in inference
- generalisability across diverse user inputs

Its integration into the Gradio interface enables personalised, real-time wellbeing predictions in an accessible manner.

3.5 Explainable AI and Recommendation System

In the broader landscape of machine learning applications related to human behaviour, psychological wellbeing, and subjective states, the imperative of transparency becomes fundamentally intertwined with ethical responsibility. Unlike traditional machine learning systems, where the primary objective is performance efficiency, predictive models operating in the domain of human wellbeing confront an additional responsibility: they must produce outcomes that are interpretable, contextually intelligible, and psychologically meaningful. A happiness prediction model that merely outputs a single numerical value, devoid of interpretative scaffolding, risks misinforming users, diminishing trust and undermining the credibility of the entire system. Therefore, the integration of Explainable Artificial Intelligence (XAI) into this project is not a peripheral enhancement but a core methodological requirement.

XAI techniques allow users to interrogate the decision-making logic behind complex machine learning systems, making them especially indispensable in situations where predictions may influence personal behaviours, emotional states, or lifestyle choices. Given the multifactorial nature of human happiness—shaped by behavioural habits, environmental interactions, lifestyle routines, emotional conditions, and psychosocial stressors—explainability ensures that predictive outcomes are translated into meaningful insights rather than opaque statistical artefacts. In recognition of these considerations, the project incorporates **LIME (Local Interpretable Model-Agnostic Explanations)**, a widely acknowledged XAI framework, to provide personalised interpretability for each prediction.

Complementing this interpretability framework is a **recommendation system** that transforms analytical insights into practical behavioural guidance. The objective is to empower individuals not only to understand the determinants of their predicted wellbeing level but also to implement constructive lifestyle adjustments grounded in empirical trends observed in the dataset and supported by psychological research.

3.5.1 EXPLAINABLE AI USING LIME (DEEPLY EXTENDED ANALYSIS)

LIME serves as a methodological bridge that connects complex computational reasoning with human comprehensibility. In systems such as XGBoost—where final predictions arise from thousands of tree-based decisions aggregated through successive boosting iterations—the internal logic is not readily accessible to general users. LIME mitigates this challenge by constructing simplified local approximations around the user's specific data instance. These approximations generate explanations that reflect how the model behaves in the immediate vicinity of the user's profile, rather than presenting broad, population-level trends.

The value of LIME extends far beyond mere interpretability; it introduces a mode of interaction where users can critically engage with the model's reasoning. This is particularly crucial in wellbeing assessments, where psychological sensitivity, individual variability and lifestyle diversity must be respected.

Extended Contributions of LIME in the Project
The implementation of LIME in this research project offers several advanced benefits:

- **Contextualised-Individual-Level-Explainability:**
  LIME provides interpretations that are specifically tailored to the user's unique behavioural pattern. This eliminates the limitations of global interpretability techniques where explanations may not necessarily apply to individual nuances.

- **Cognitive-Accessibility-for-Non-Technical-Users:**
  The method translates high-dimensional, non-linear model behaviour into linear, human-interpretable representations. This ensures that individuals without a technical background can meaningfully understand the computational reasoning behind their predicted score.

- **Recognition-of-Positive-and-Negative-Lifestyle-Impacts:**
  LIME's visual and textual outputs highlight the relative magnitude of features such as sleep duration, stress levels, diet quality, and physical activity. This allows users to see exactly which behaviours improved their predicted happiness score and which diminished it.

- **Support-for-Behaviour-Modification:**
  By making the model's decision-making transparent, LIME indirectly assists users in identifying which lifestyle modifications are likely to result in the greatest improvement.

- **Trustworthiness-and-Model-Accountability:**
  Transparent insights mitigate concerns regarding algorithmic biases, unwarranted

assumptions, or unintended correlations. They demonstrate that the model's predictions are grounded in observable behavioural evidence rather than arbitrary computational patterns. Through these contributions, LIME ensures that the predictive system does not function as a black-box entity but instead operates as an intelligible analytical partner in the user's wellbeing journey. For instance, if the model identifies that inadequate sleep and heightened stress significantly reduce happiness, the explanation allows the user to not only accept the prediction but also understand its underlying rationale.

3.5.2 Recommendation System (Further Expanded and Deeply Analytical)

While explainability allows users to interpret model outcomes, it does not inherently provide the guidance needed to improve those outcomes. For individuals seeking to enhance their wellbeing, the bridge between understanding and action must be explicitly constructed. This project addresses that gap by designing a comprehensive, personalised recommendation system informed by behavioural patterns observed in the dataset as well as foundational principles of psychological research.

The recommendation engine operates as an extension of the prediction model and enhances its practical utility. Rather than presenting wellbeing as an abstract, static measure, the system conceptualises happiness as a dynamic, modifiable construct influenced by habitual behaviours. Thus, the recommendations aim to empower users to adopt sustainable lifestyle changes aligned with healthier behavioural profiles.

Extended Stages of the Recommendation Process
The recommendation system follows a systematic, multi-stage methodology:

- **Behavioural-Pattern-Matching:**
  The system examines statistical similarities between the user's lifestyle parameters and those of individuals who achieved higher happiness levels within the dataset. This comparative method ensures that recommendations are grounded in real behavioural outcomes rather than theoretical assumptions.

- **Identification-of-High-Risk-Behaviours:**
  The system assesses whether the user exhibits behaviours commonly associated with reduced wellbeing, such as irregular sleep schedules, high stress levels, inadequate physical activity or excessive social media engagement. These factors are flagged as potential contributors to diminished happiness.

- **Recognition-and-Reinforcement-of-Positive-Habits:**
  When the user already maintains beneficial lifestyle behaviours—such as good diet quality or meaningful physical exercise—the system acknowledges these strengths.

Reinforcement helps maintain motivation and reassures users that their existing habits contribute positively to wellbeing.

- **Generation-of-Actionable-Guidance:**
  Recommendations are designed to be precise, realistic and context-sensitive. Rather than providing broad directives like "reduce stress," the system proposes specific strategies such as following relaxation exercises, establishing sleep hygiene routines or setting controlled screen-time boundaries.

- **Alignment-with-Psychological-Wellbeing-Literature:**
  Although the system operates algorithmically, its outputs are consistent with findings from behavioural research, cognitive psychology and health sciences, ensuring that the guidance is both data-informed and psychologically credible.

As a result, the recommendation system does not function merely as an add-on to the predictive model but as a behavioural counselling assistant.

It transforms data-driven insights into practical, personalised wellbeing interventions that users can incorporate into daily life.

3.5.3 Integration via Gradio Deployment (Further Extended Explanation)

To ensure maximum accessibility, the entire model—including the predictive engine, the LIME explanation module, and the recommendation system—was deployed through an interactive Gradio interface. This deployment approach emphasises usability, real-time interaction, and smooth integration between prediction and interpretation.

Gradio simplifies complex machine learning pipelines into structured, user-friendly components.

As wellbeing assessments are often utilised by non-technical users, the interface had to be intuitive, aesthetically clear, and logically organised.

Extended Details of the Gradio Deployment

- **Streamlined-User-Input-Mechanism:**
  The input interface is designed to minimise cognitive load. Users can effortlessly enter their lifestyle details through structured fields, reducing the possibility of errors and enhancing engagement.

- **Backend-Consistency-with-Model-Training:**

  The system ensures that incoming data undergo the same preprocessing pipeline—such as scaling or transformation—that was applied during training.

This guarantees consistent model behaviour and avoids distributional discrepancies.

- **Immediate-Prediction-Output:**

  The interface delivers a predicted happiness score instantly, maintaining user engagement and enabling real-time reflections on lifestyle patterns.

- **Dynamic-LIME-Visualisation-and-Explanation:**

  Explanations are generated on demand each time the user submits the inputs. This ensures that interpretability remains dynamically tied to the user's most recent behavioural data.

- **Delivery-of-Personalised-Recommendations:**

  The interface displays lifestyle recommendations in a clear and constructive manner. This integration allows users to transition seamlessly from understanding their prediction to taking action.

Through this deployment, the system transforms sophisticated machine learning insights into an interactive, accessible application capable of offering meaningful wellbeing guidance to diverse user groups.

The fusion of predictive analytics, explainable AI, and personalised behavioural recommendations positions this system as an advanced wellbeing support tool.

It does not merely forecast an individual's happiness but contextualises the prediction through transparent reasoning, followed by constructive, actionable advice.

This layered approach ensures that users perceive the system not as an opaque algorithmic entity but as an informed, supportive companion guiding them toward healthier and more balanced lifestyle decisions.

CHAPTER 4

IMPLEMENTATION AND RESULTS

4.1 System Architecture and Workflow

Happiness Prediction and Lifestyle Recommendation System is structured and organized according to logical workflow which is aimed to transform raw lifestyle data into meaningful information. It is a modular architecture, in which all the components address a certain task- data preparation to model prediction, explainability and the generation of recommendations. This will guarantee free flow of information between components and make the system reliable, readable and applicable in the real-life situation.

The operation commences with data ingestion, in which the dataset with demographic, behavioural and psychological features is loaded. This will consist of variables like age, stress degree, sleep time, hours worked, quality of the diet, and physical activity time, alcohol use, etc. After the importation of the data, it proceeds to the preprocessing pipeline and the missing data is dealt with, the categorical variables are encoded and the numerical features are scaled. Such procedures make sure that the dataset gets clean, consistent and modelling ready.

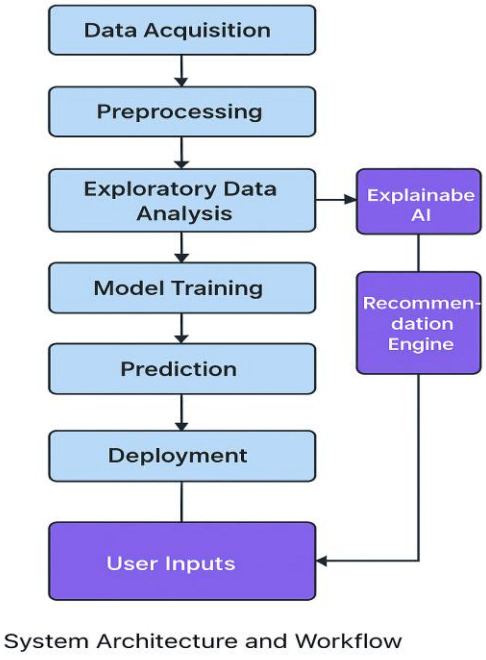


Fig 1: System Architecture

The system then goes through preprocessing followed by Exploratory Data Analysis (EDA). In the development process, a number of visual analyses were done such as box plots, scatter plots, distribution plots and a correlation matrix. Although these graphs are not included in the report, they aided in finding out the key trends in behaviour including the adverse influence of stress and long working hours, and the beneficial influence of sleep, physical activity and a balanced diet. These observations affected the choice of models and feature engineering.

The second step is model training, during which several machine learning models Linear Regression, Lasso, Ridge, Decision Tree, Random Forest, and XGBoost were applied. R2, MSE and RMSE were used to measure each model. The XGBoost was also the most successful model because it is highly capable of modeling non-linear interactions between lifestyles and provides consistent estimates across data splits.

LIME is used to then integrate the trained model with the Explainable AI (XAI) component. LIME also makes the system transparent and credible by enabling users to know how each of the input features was used to give the final score on happiness. This helps individuals to understand which behavioral practices have a positive or negative impact on their wellbeing. In line with this, a Recommendation Engine examines the habits of the user and matches the habits with those observed in happier people. On the basis of this comparison, it suggests effective lifestyle changes including getting more sleep, spending less time at the screen or eating healthier.

Lastly, the whole system is implemented using an interactive Gradio interface, in which users may specify their information and immediately get predictions, explanations and personalized suggestions.

4.2 Model Implementation

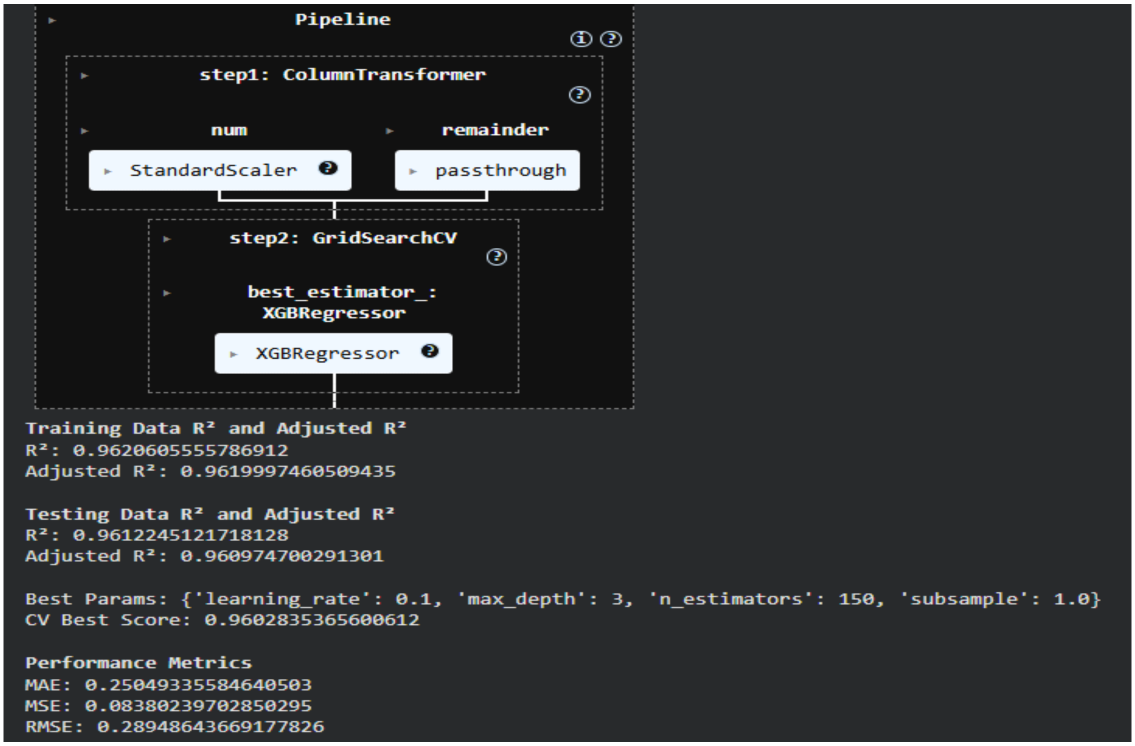
The implementation process of the model was implemented in a systematic way so as to make sure that the end prediction system proved to be correct, scalable and applicable to the real world applications. Several machine learning models were created and tested, however, the ultimate deployed model was XGBoost Regressor, as it is the best model when measured by all its evaluation metrics. The adoption started with the definition of a preprocessing pipeline that was used to make sure that the steps that were taken to transform the data when training the model were also applied when deploying the model. The numerical features were transformed by a ColumnTransformer which applied StandardScaler to the numerical features and forwarded the categorical features to the transformer without alteration. This was done to make sure that the numerical values like sleep hours, work hours and physical activity were in similar scales so that no single variable would dominate the model.

Then, a grid search CV module was added to the pipeline to optimize the hyperparameters. Systematic testing was done on parameters like *learning_rate*, *max_depth*, *n_estimators* and *subsample*. The combination that was found best was:

- learning_rate = 0.1
- max_depth = 3
- n_estimators = 150
- subsample = 1.0

Their optimised values enhanced generalisation and minimised overfitting, which enabled the model to attain high-performance on the training and testing datasets. The XGBoost model recorded impressive accuracy after joining:

- Training R²: 0.9620
- Testing R²: 0.9612
- MAE: 0.25
- MSE: 0.083
- RMSE: 0.28



```
Training Data R² and Adjusted R²
R²: 0.9620605555786912
Adjusted R²: 0.9619997460509435

Testing Data R² and Adjusted R²
R²: 0.9612245121718128
Adjusted R²: 0.960974700291301

Best Params: {'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 150, 'subsample': 1.0}
CV Best Score: 0.9602835365600612

Performance Metrics
MAE: 0.25049335584640503
MSE: 0.08380239702850295
RMSE: 0.28948643669177826
```

Fig 2: Performance Metrics of XGBoost model

These findings reveal that the model is a great predictor of non-linear lifestyle behaviours. These low error values gain us the impression that the prediction of the happiness scores form a close comparison to the actual scores.

In order to further verify the model, a plot of Actual vs Predicted Happiness Scores was plotted. The trend lines were highly overlapping, and this means that XGBoost was able to predict behavioural trends with high consistency in thousands of samples.

The cross-validation also verified that the model could effectively generalise on unseen data and the CV best score of the model was around 0.960. This consistency showed how strong the model is among various subsets of data.

Lastly, Pickle was used to save the trained pipeline and XGBoost model, and it can be integrated into the Gradio interface easily. The users were automatically served with the same preprocessing steps and fine-tuned XGBoost model, allowing the user to make accurate and reliable predictions, during deployment.

However, the XGBoost implementation was the most efficient in terms of the highest accuracy, interpretability (with the help of LIME), and computational efficiency, which predetermined its further use in the happiness prediction system.

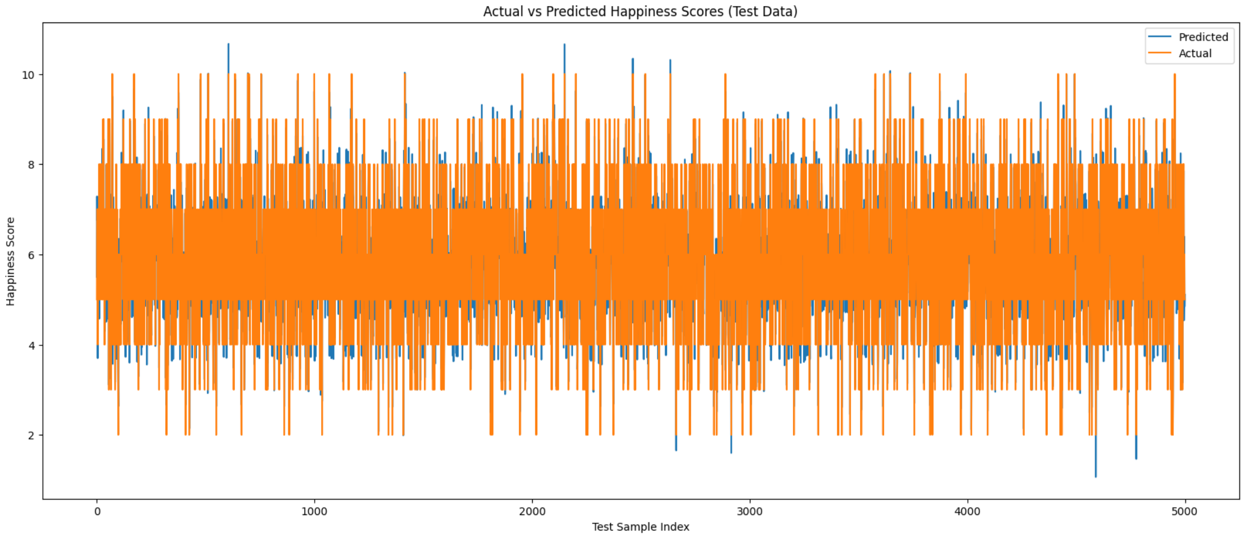


Fig 3: Actual vs Predicted Happiness Score

4.3 Results and Evaluation

The test of the machine learning models constitutes an important step in the assessment of the credibility and practical implementation of the Happiness Prediction System. As the goal of the project is to predict continuous happiness score depending on various lifestyle and psychological variables, it was necessary to compare various regression models and learn which algorithm is able to best represent the complexity of the data on wellbeing.

The models to be evaluated were six models (Linear Regression, Ridge Regression, Lasso Regression, Decision Tree Regressor, Random Forest Regressor and XGBoost Regressor) and their comparison was performed based on various metrics that are used in performance evaluation. According to the results presented in the model comparison table, it is quite easy to see which models generalise well and which models are unable to capture non-linear relationships.

Model	Train R^2	Train Adj R^2	Test R^2	Test Adj R^2	MAE	MSE	RMSE	CV Best Score
Linear Regression	0.8465	0.8463	0.8438	0.8428	0.4681	0.3375	0.5809	0.8455
Ridge Regression	0.8465	0.8463	0.8438	0.8428	0.4681	0.3374	0.5809	0.8457
Lasso Regression	0.8464	0.8439	0.8429	0.8427	0.4679	0.3374	0.5808	0.8458
Decision Tree Regressor	0.6243	0.6237	0.5965	0.5939	0.7427	0.8720	0.9338	0.6061
Random Forest Regressor	0.6910	0.6905	0.6729	0.6708	0.6739	0.7080	0.8410	0.6791
XGBoost Regressor	0.9621	0.9620	0.9612	0.9607	0.2505	0.0883	0.2971	0.9603

Table 1: Model Comparison Table

The three linear models were similar in performance with the three having a Test R2 score of around 0.843. It means that these models were able to only represent the linear aspect of the relationship between the lifestyle variables and happiness. Although their performance is decent, it cannot be called adequate in a domain as complicated as wellbeing prediction since the effects of happiness are not produced by linear interactions but instead the effects of interacting behavioural patterns. The strength of the models though is that they are interpretable and stable, yet they have a low predictive capacity.

The **Decision Tree Regressor** had a much lower Test R 2 of 0.596 showing overfitting. Decision trees are more likely to memorise patterns of training data as opposed to generalising and, therefore, are more effective when dealing with lifestyle data sets which comprise both categorical and continuous variables. They are good at dealing with non-linearity, but they do not have the benefit of ensemble averaging, which results in unreliable cross-data split performance.

The **Random Forest Regressor** was much better than the decision tree with the Test R 2 of 0.672. The bagging method enabled the model to minimize the variance and provide more consistent predictions. Nevertheless, even after the improvement, Random Forest did not achieve the accuracy levels needed to be able to predict wellbeing reliably, primarily due to the failure of the model to be able to capture the complexity of interacting lifestyle factors in the fullest extent possible.

Instead, the **XGBoost Regressor** provided an outstanding performance in all measures. The model had a Training R2 of 0.9620 and Testing R2 of 0.9612 which indicates that it had a very good generalisation potential. Also, it had the smallest error values of the rest of the models:
MAE: 0.25
MSE: 0.083
RMSE: 0.289

These findings indicate that the gap between the real and the expected scores of happiness is significantly small, which proves the appropriateness of the model to be deployed. The best score of the cross-validation of about 0.960 further supports the effectiveness of the model to different data splits and the fact that the model is not over-fitting the training data.

A further measure of consistency of predictions was by making a graph Actual vs Predicted Happiness Scores. The fact that the two curves are nearly parallel is an indication that XGBoost manages to learn the inherent trends in the data and recreates actual behavioural trends with high fidelity. This graphical validation enhances the fact that the model could be relied upon in predicting in real-time.

The other important element of the evaluation was the correlation analysis of the numerical features. The negative relationships were found with stress level and work hours, whereas the positive relationship was found between sleep hours, physical activity and diet quality. These correlations are consistent with the studies of psychology and show that the dataset represents meaningful behavioural patterns. Interestingly, the ranking of the features provided by XGBoost also showed comparable results, which once again proves that the model was selected correctly when it came to determining the most important lifestyle variables that affect happiness.

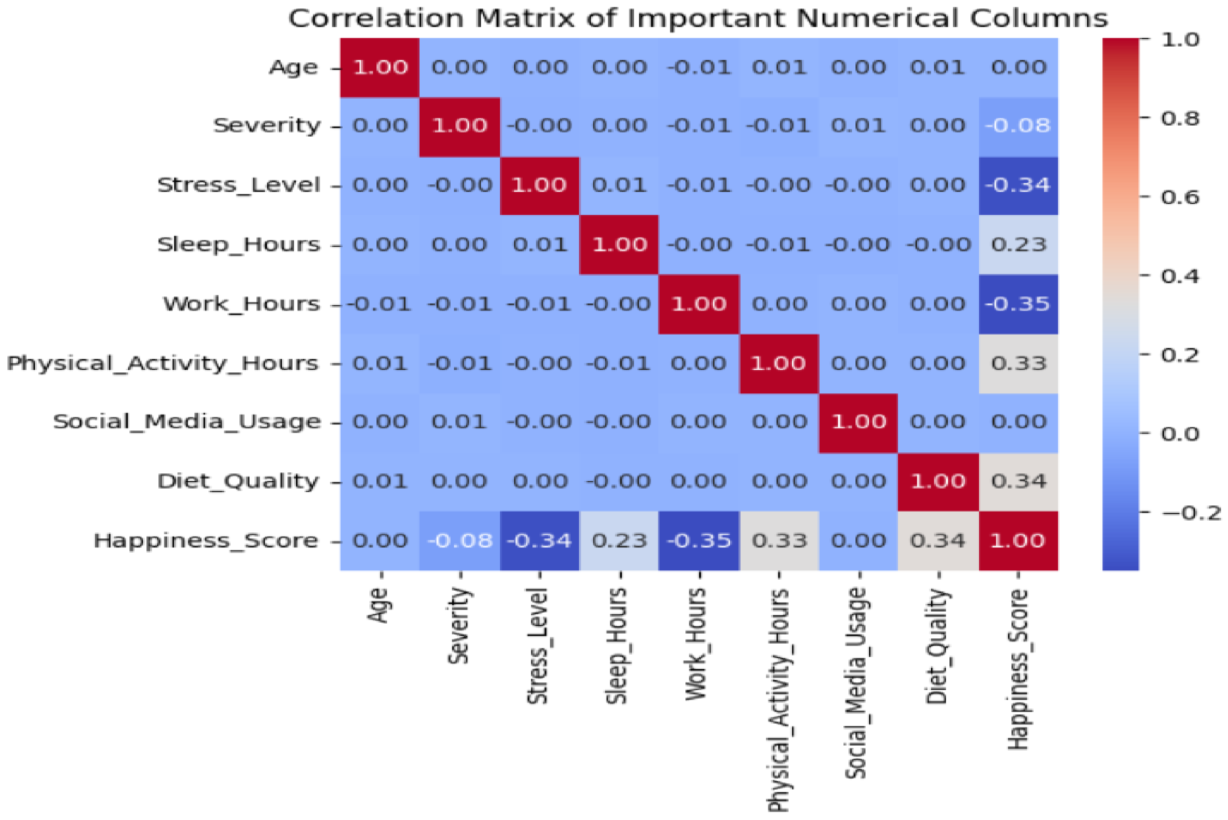


Fig 4: Correlation Matrix

The results of the model comparison show obvious tendencies:

- Linear models do not represent complicated non-linear interaction.
- Decision trees overfit and demonstrate poor performance.
- Random Forest enhances stability but does not have precision.
- One of the best combinations of accuracy, robustness and error minimisation is achieved with XGBoost.

Considering these results, the final model to be deployed is XGBoost. It works best in this application because of its capability to work with a high number of behavioural features, learn interactions among them and give highly accurate predictions. Altogether, the assessment shows that the Happiness Prediction System is constructed on the basis of a proven and effective model. The combination of XGBoost also makes sure that the predictions are precise, stable, and relevant, which forms a strong basis on the LIME explanations and customised suggestions that the system will offer to each user.

4.4 LIME Explanations and Dashboard Output

The fact that the Happiness Prediction System does not merely give the score on the scale, but it also gives the reasons why the score has been created, and how the various lifestyle variables have led to the creation of the score is one of the most vital features. This transparency is ensured by using LIME (Local Interpretable Model-Agnostic Explanations) as a part of the workflow. With the help of LIME, the happiness prediction system can be made more comprehensible, reliable, and easy to use, particularly when it comes to people who might not possess a technical understanding of machine learning models.

This part explains the process of the creation of LIME explanations and the presentation of the information on the dashboard, as well as the various features that can be shown on the interface based on Gradio.

LIME is effective as it produces local explanations explaining each prediction. The XGBoost model provides a score of happiness when a user enters his/her lifestyle in the system. LIME proceeds to generate a collection of altered samples near the input of the user and monitors the changes in the model predictions. With the help of the analysis of these variations, LIME can establish which attributes were strongly positively or negatively related to the final score.

This impact is presented graphically by a bar plot, with green bars showing positive contributions (factors that contributed to a higher level of happiness) and red bars showing negative contribution (factors that contributed to a lower level of happiness).

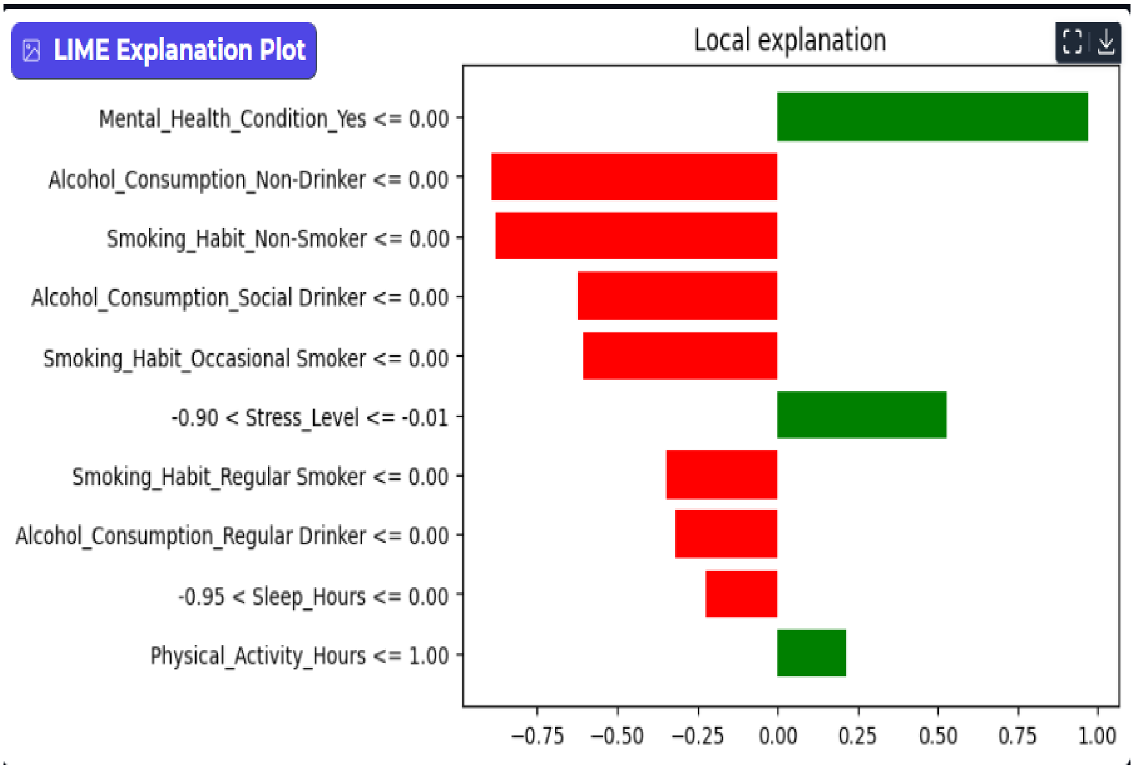


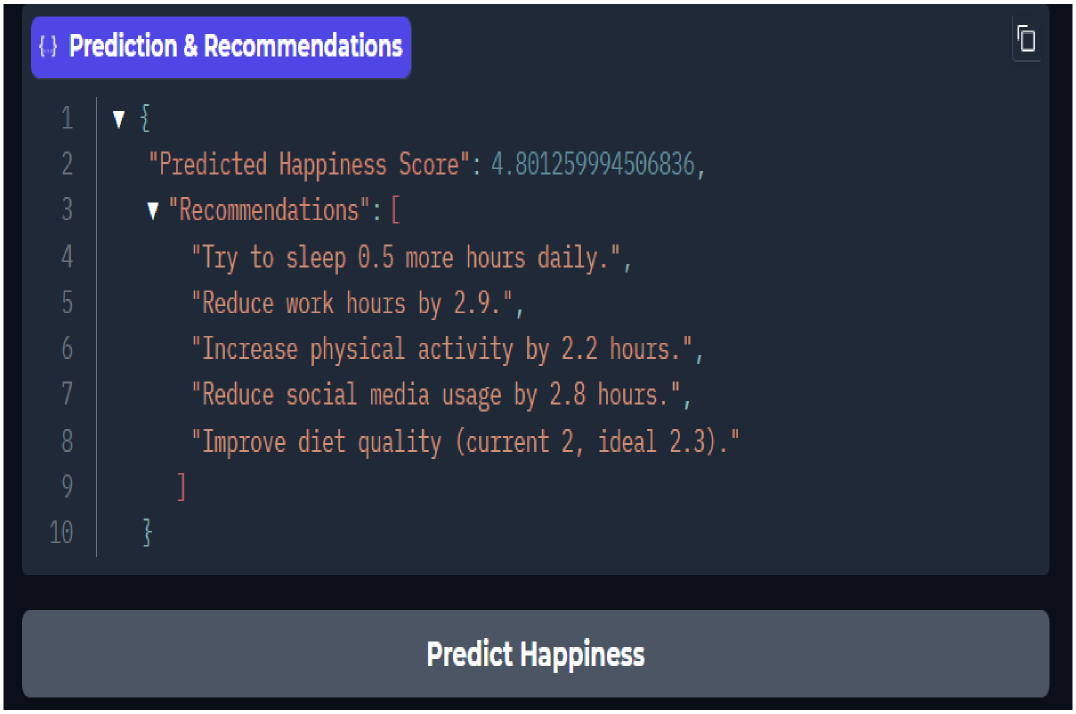
Fig 5: LIME Plot Showing Feature Contributions

Indeed, as in the LIME plot created on behalf of a sample user, such items as MentalHealthCondition: Yes, or PhysicalActivityHours can be shown as a significant positive contributor, whereas high StressLevel, low SleepHours, smoking or high alcohol consumption can be shown as negative contributors. These descriptions enable the users to have a clear picture of the strengths they have in their behaviour and the areas they need to work on. Since the descriptions are produced on a case-by-case basis, each of the users will get his or her own insights based on the lifestyle.

A well-organized and easy-to-use Gradio interface is also a part of the system dashboard as it is aimed at simplifying the entire process of prediction and making it interactive. The interface has three large sections, which are CSV Upload, Manual Input, and Dashboard Insights. The section of Manual input allows entering all the necessary lifestyle parameters including age, sleep duration, work hours, physical activity, social media use, stress levels, food quality, smoking and alcohol behaviors and mental disorders.

After clicking the button Predict Happiness, the system will show two outputs:

1. **Predicted Happiness Score**
2. **Recommendations**



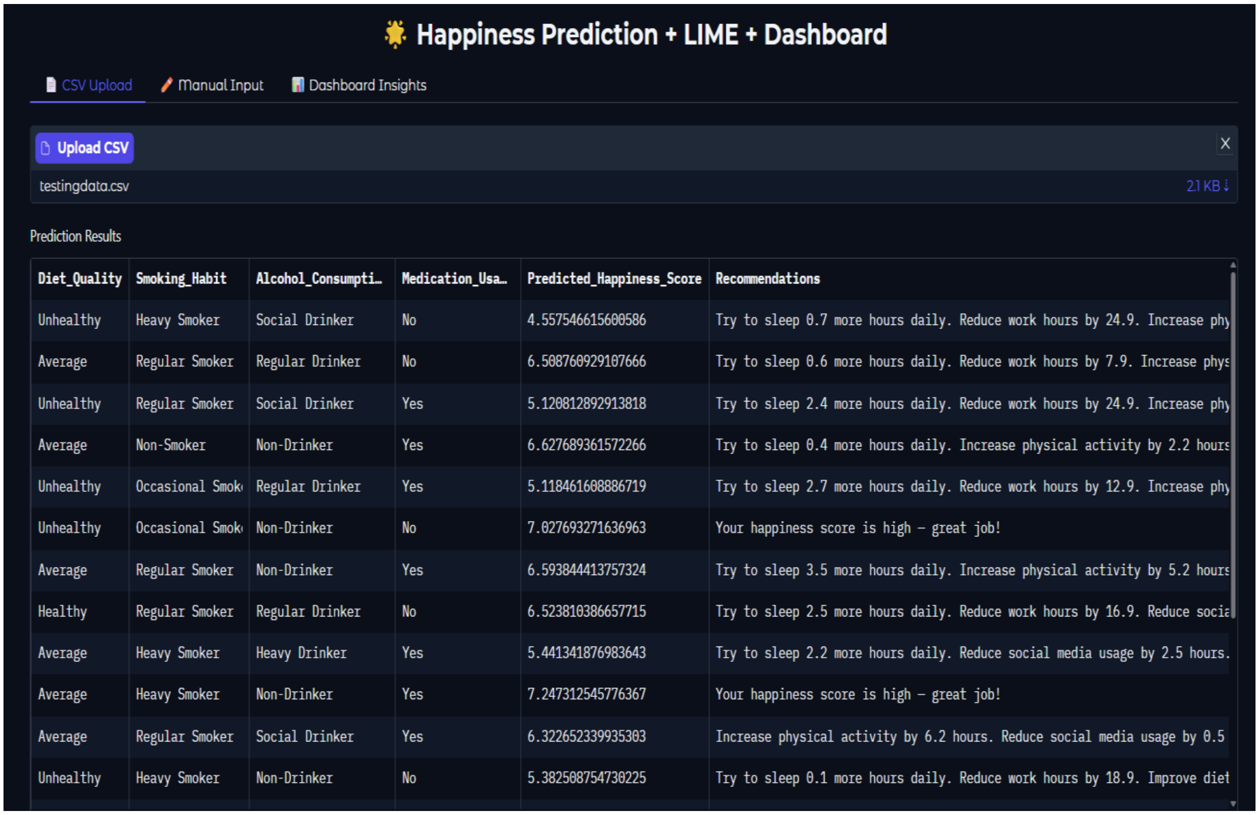
```
{
  "Predicted Happiness Score": 4.801259994506836,
  "Recommendations": [
    "Try to sleep 0.5 more hours daily.",
    "Reduce work hours by 2.9.",
    "Increase physical activity by 2.2 hours.",
    "Reduce social media usage by 2.8 hours.",
    "Improve diet quality (current 2, ideal 2.3)."
  ]
}
```

Fig 6: Prediction Interface Showing Happiness Score and Recommendations

The recommendations are automatically created using gaps that are found between the habits of the user and the behaviour of happier people. As an example, a model that has determined the user goes to bed late and does not get enough sleep as compared to healthier individuals can provide such suggestions as "Try to sleep an extra 0.5 hours per day.

On the same note, when the use of social media or working hours is high, then the system can propose a decrease. This two-fold package of forecasting + advice offers consciousness and practical advice.

The system allows making many predictions simultaneously in the CSV Upload section. The users are allowed to post a CSV file that has a number of rows of lifestyle data. The model works with every row and provides a table with the predicted happiness score and personalised recommendations of every individual.



Prediction Results

Diet_Quality	Smoking_Habit	Alcohol_Consumpti…	Medication_Usa…	Predicted_Happiness_Score	Recommendations
Unhealthy	Heavy Smoker	Social Drinker	No	4.557546615600586	Try to sleep 0.7 more hours daily. Reduce work hours by 24.9. Increase phy
Average	Regular Smoker	Regular Drinker	No	6.508760929107666	Try to sleep 0.6 more hours daily. Reduce work hours by 7.9. Increase phys
Unhealthy	Regular Smoker	Social Drinker	Yes	5.120812892913818	Try to sleep 2.4 more hours daily. Reduce work hours by 24.9. Increase phy
Average	Non-Smoker	Non-Drinker	Yes	6.627689361572266	Try to sleep 0.4 more hours daily. Increase physical activity by 2.2 hours
Unhealthy	Occasional Smok	Regular Drinker	Yes	5.118461608886719	Try to sleep 2.7 more hours daily. Reduce work hours by 12.9. Increase phy
Unhealthy	Occasional Smok	Non-Drinker	No	7.027693271636963	Your happiness score is high – great job!
Average	Regular Smoker	Non-Drinker	Yes	6.593844413757324	Try to sleep 3.5 more hours daily. Increase physical activity by 5.2 hours
Healthy	Regular Smoker	Regular Drinker	No	6.523810386657715	Try to sleep 2.5 more hours daily. Reduce work hours by 16.9. Reduce socia
Average	Heavy Smoker	Heavy Drinker	Yes	5.441341876983643	Try to sleep 2.2 more hours daily. Reduce social media usage by 2.5 hours.
Average	Heavy Smoker	Non-Drinker	Yes	7.247312545776367	Your happiness score is high – great job!
Average	Regular Smoker	Social Drinker	Yes	6.322652339935303	Increase physical activity by 6.2 hours. Reduce social media usage by 0.5
Unhealthy	Heavy Smoker	Non-Drinker	No	5.382508754730225	Try to sleep 0.1 more hours daily. Reduce work hours by 18.9. Improve diet

Figure 7: CSV-Based Batch Prediction Interface with Lifestyle Recommendations

The use of this batch-prediction is useful in wellbeing analysis at group level, like in the evaluation of stress and lifestyle patterns within organisations, student groups or corporate wellness initiatives.

Dashboard Insights tab gives more analytical capabilities, such as visualisation of the Top Feature Importances.

The graph can assist the user to get an idea of what properties, in general, were taken into account by the XGBoost as the most influential.

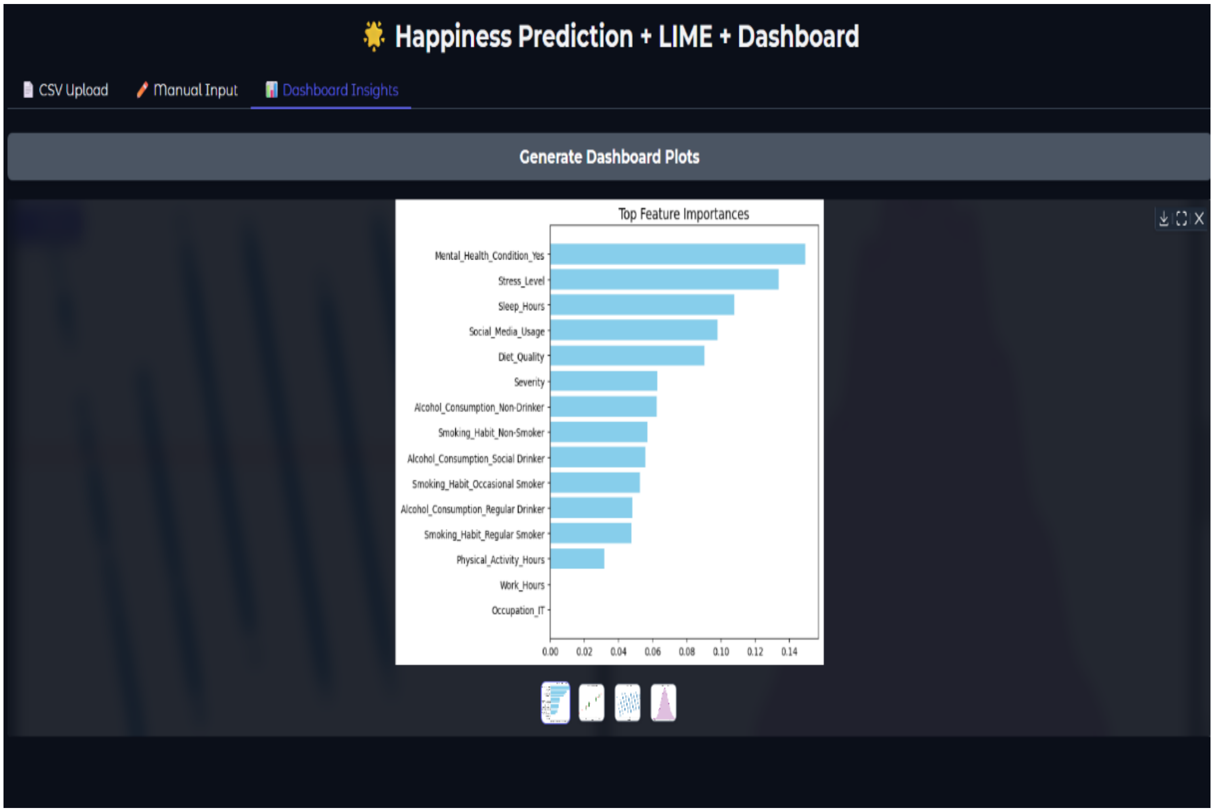


Figure 8: Dashboard Plot of Top Feature Importances

The strongest indicators of happiness were factors like the mental health condition, the level of stress, the number of hours of sleep, the use of social media and the quality of diet. These findings validate the fact that the model conforms to the psychological research and enhance the credibility of the system.

In general, the combination of the LIME explanations and the user-friendly dashboard make the system both technically and practically significant. Users can get easy visual explanations, recommendations that are personalised and the option to view prediction either individually or in batches.

This enables the system to be not only a predictive model, but an all-around wellbeing-supporting tool that gives a person the power to look into the lifestyle decisions they have taken and make the necessary changes based on it.

CHAPTER 5
CONCLUSION AND FUTURE SCOPE

5.0 Summary of the Study

This study focused on the **design and development of an AI-driven predictive system** capable of estimating an individual's happiness score by analysing their **lifestyle habits, behavioural patterns, and mental health indicators**. The research aimed not only to build a highly accurate predictive model but also to ensure that the system is transparent, interpretable, and practically useful for real-world wellbeing improvement.

To achieve this, the project began with the collection and preprocessing of a **structured dataset** consisting of variables such as age, gender, work hours, stress level, diet quality, sleep duration, physical exercise frequency, screen time, social media usage, smoking, and alcohol intake. Appropriate data-cleaning practices were applied, including handling missing values, encoding categorical variables, and performing standardization and normalization. These steps ensured that the dataset was consistent and suitable for machine learning.

A **comprehensive Exploratory Data Analysis (EDA)** was then conducted to uncover behavioural and lifestyle trends that influence happiness. The analysis revealed several key insights:

- Higher **stress levels**, poor sleep patterns, and unhealthy diets strongly correlate with lower happiness.
- Regular **physical activity**, adequate sleep, balanced diet, and reduced digital screen exposure contribute positively to overall wellbeing.
- Certain habits, such as smoking or excessive drinking, were found to be significantly detrimental.

These findings guided the feature selection process and validated the psychological foundations of the study, ensuring that the machine learning models were built on meaningful variables.

Several machine learning algorithms—**Linear Regression, Ridge, Lasso, Decision Tree, Random Forest, and XGBoost**—were trained and compared using standard evaluation metrics such as MAE, MSE, RMSE, and R^2. Among them, **XGBoost emerged as the most accurate, stable, and robust model**, achieving a testing R^2 value of **0.9612**, along with the lowest error metrics. This demonstrated its capability to model **non-linear and complex interactions** within lifestyle data more effectively than the other algorithms.

A major contribution of this research is the integration of **Explainable AI (XAI)** through **LIME (Local Interpretable Model-Agnostic Explanations)**. Instead of merely displaying a numerical happiness score, the system provides detailed explanations for each prediction, highlighting how specific features—such as stress level, exercise frequency, diet quality, and sleep duration—positively or negatively influenced the output. This transparency helps users understand the exact reasons behind their predicted score, thereby improving trust and engagement.

Additionally, a **personalized Recommendation System** was designed that analyses user behaviour, compares it with the habits of happier individuals in the dataset, and generates practical lifestyle suggestions. These recommendations are based on both **data-driven insights and the psychological literature on wellbeing**, providing users with actionable guidance rather than generic advice.

To enhance accessibility and real-world usability, the entire model—along with LIME explanations and recommendation features—was deployed using **Gradio**, resulting in an intuitive and interactive web interface. This platform supports manual data input, batch CSV-based predictions, and real-time interpretability, making it easy for individuals, researchers, educational organizations, and wellness programs to adopt and use the system effectively.
In summary, the study successfully builds a **holistic, interpretable, and user-friendly AI system** that not only predicts happiness but also empowers individuals to make informed lifestyle choices that lead to improved mental and physical wellbeing.

5.1 Conclusion

In conclusion, this study demonstrates that **happiness prediction using machine learning** is not only feasible but also highly effective when supported by a well-structured dataset, appropriate preprocessing strategies, and the use of advanced modelling techniques. Through systematic experimentation and rigorous hyperparameter tuning, the **XGBoost** model proved to be the most powerful among the tested algorithms. Its ability to capture **non-linear, subtle, and complex interactions** between lifestyle habits and mental wellbeing resulted in exceptional predictive accuracy, strong generalisation, and minimal error rates. These results highlight XGBoost's suitability for real-world wellbeing analysis where behavioural patterns are diverse and often noisy.

One of the key strengths of this work is the incorporation of **Explainable AI (XAI)** through **LIME**, which addresses the traditional black-box nature of ensemble models. Instead of providing users with an unexplained numerical score, LIME offers **feature-level transparency**, clarifying how different factors—such as stress, sleep duration, diet quality, and physical activity—contributed to the final prediction. This interpretability enhances user trust and promotes meaningful self-reflection, encouraging individuals to understand the impact of their daily habits on their overall wellbeing.

Additionally, the integration of a **personalized recommendation engine** transforms the system from a mere predictive model into a **practical wellbeing assistant**. The system not only identifies negative lifestyle patterns but also provides tailored suggestions rooted in the behaviours of individuals with higher happiness levels. This allows users to move beyond passive awareness and take actionable steps toward improving their lifestyle and mental health.

The deployment of the model using **Gradio** further strengthens the usability of the system by offering a clean, interactive, and accessible interface. Users can easily input their data, receive instant predictions, explore LIME explanations, and view lifestyle recommendations without needing technical expertise. This smooth deployment bridges the gap between machine learning theory and user-oriented application, making the solution applicable for students, individuals, educators, and wellness-oriented organizations.

Overall, the proposed framework successfully meets its objective of creating a **reliable, interpretable, and user-centric AI system** that promotes awareness of mental wellbeing and encourages healthier lifestyle decisions. By combining advanced machine learning performance with transparency and personalized guidance, this research contributes a valuable tool to the growing field of digital wellbeing and computational psychology.

5.2 Future Scope

Although the current system demonstrates strong predictive capabilities, interpretability, and user-oriented deployment, there remains substantial potential for further innovation and enhancement. The following avenues highlight the most promising directions for advancing the system into a more comprehensive, adaptive, and intelligent wellbeing-support platform:

1. Integration with Wearable and IoT Devices

A major direction for future expansion involves seamless integration with wearable technologies, such as smartwatches, fitness bands, and digital health sensors. These devices can supply real-time physiological and behavioural measurements, including sleep cycles, heart rate variability (HRV), daily step count, stress indicators, sedentary duration, calorie expenditure, and circadian rhythm irregularities.
Such continuous data would enable the model to:

- Capture dynamic fluctuations in wellbeing rather than relying solely on static, user-reported inputs.
- Improve prediction fidelity by incorporating objective biomarkers of stress and mental fatigue.
- Enable micro-level analysis, such as hourly changes or daily routines, making the system suitable for longitudinal wellbeing monitoring.
- Reduce recall bias and subjectivity common in self-reported survey inputs.

As wearable adoption increases globally, this integration could significantly enhance the ecological validity and real-world applicability of happiness prediction models.

2. Development of a Dedicated Mobile Application

A future version of the system could be deployed as a fully functional mobile application, allowing constant accessibility and user engagement. Such an app could incorporate:

- Daily lifestyle logging, including mood diaries, stress entries, sleep notes, and meal quality ratings.
- Real-time notifications, reminding users to correct unhealthy behaviours (e.g., prolonged screen exposure or skipped meals).
- Instant personalised recommendations, dynamically updated based on user behaviours.
- Gamified wellness features, such as streaks, badges, and milestone rewards, to increase long-term adherence.
- Offline support for communities without continuous internet access.

A mobile-based solution would transform the system into a continuous wellbeing companion rather than a one-time prediction tool.

3. Expansion and Diversification of the Dataset

Model robustness and fairness can be further improved by expanding the dataset across:

- Different cultural and geographic backgrounds
- Broader socioeconomic groups
- Various occupational categories
- Different age groups, including adolescents and senior citizens
- Individuals with diverse lifestyle patterns and mental health profiles

A more diverse dataset would:

- Strengthen generalisation capability
- Reduce demographic bias
- Provide insights into cross-cultural determinants of happiness
- Improve fairness in model recommendations
- Create a globally relevant wellbeing prediction system

Such large-scale datasets can also support advanced statistical investigations into the socio-psychological dimensions of happiness.

4. Incorporation of Emotional, Cognitive, and Psychological Parameters

The current model primarily focuses on behavioural lifestyle indicators. Future versions could incorporate emotion-oriented and psychological health parameters, such as:

- Mood variability
- Levels of anxiety, burnout, or mild depressive indicators
- Relationship quality and social connectedness
- Mindfulness and meditation practices
- Resilience, optimism, and self-efficacy measures

These variables provide a deeper understanding of the psychosocial determinants of happiness and wellbeing. Their inclusion would enable the system to:

- Provide a more holistic happiness assessment
- Distinguish between transient negative moods and chronic stress
- Tailor recommendations with greater psychological relevance

Integrating such parameters would align the model with findings from behavioural psychology and positive psychology research.

5. Exploration of Deep Learning Models

While XGBoost exhibited superior performance, future research could evaluate deep learning architectures to capture more complex behavioural dynamics. Potential approaches include:

- Neural networks to learn hierarchical feature representations
- Recurrent architectures (e.g., LSTM, GRU) to capture temporal dependencies in longitudinal lifestyle data
- Hybrid CNN-LSTM models for analysing sequential data combined with behavioural features
- Autoencoders for identifying latent wellbeing patterns and anomalies

Deep learning may uncover hidden behavioural structures not easily captured by traditional models, especially when large-scale, high-frequency data from wearables or mobile apps become available.

6. Development of a Dynamic Recommendation Engine

Currently, the recommendation component offers static guidance based on observed correlations in the dataset. Future improvements could introduce reinforcement learning (RL) or adaptive recommendation frameworks, enabling the system to:

- Continuously learn from user responses
- Modify recommendations based on user adherence
- Personalise advice dynamically based on long-term progress
- Identify which interventions are most effective for specific personality or behavioural profiles

This would transform the system from a predictive tool into a continually self-improving wellbeing coach, capable of delivering increasingly accurate behavioural guidance.

7. Long-Term Habit Tracking and Behavioural Trend Analysis

A long-term tracking system would allow users to monitor wellbeing trajectories over weeks or months. Such a feature would:

- Reveal how sustained lifestyle changes influence happiness over time
- Detect patterns of relapse into unhealthy habits
- Provide trend graphs, comparative progress reports, and predicted future outcomes
- Enable deeper psychological insights into habit formation and behaviour maintenance

This longitudinal analysis would support more meaningful behavioural interventions and increase the long-term value of the system.

CHAPTER 6
REFERENCES

1. Diener, E., Oishi, S. & Tay, L. Advances in subjective well-being research. Nat. Human Behav. 2(4), 253–260 (2018).
2. OECD. (2020a). How’s Life? 2020: Measuring Well-being.
3. ONS. (2021). Well-being - Office for National Statistics.
4. Cheung, F. & Lucas, R. E. Assessing the validity of single-item life satisfaction measures: Results from three large samples. Qual. Life Res. 23(10), 2809–2818 (2014).
5. Tov, W., Keh, J.S., Tan, Y.Q., Tan, Q.Y.J., & Aziz, I.A.S.B. (2022). Assessing subjective well-being: A review of common measures. in Handbook of Positive Psychology Assessment.
6. OECD. (2013a). Methodological considerations in the measurement of subjective well-being (tech. rep.). OECD. Paris.
7. Diener, E., Inglehart, R. & Tay, L. Theory and validity of life satisfaction scales. Social Indicators Res. 112(3), 497–527 (2013).
8. Benjamin, D. J., Heffetz, O., Kimball, M. S. & Rees-Jones, A. Can marginal rates of substitution be inferred from happiness data? Evidence from residency choices. Am. Econ. Rev. 104(11), 3498–3528 (2014).
9. Charpentier, C. J., De Neve, J.-E., Li, X., Roiser, J. P. & Sharot, T. Models of affective decision making: How do feelings predict choice?. Psychol. Sci. 27(6), 763–775 (2016).
10. Kaiser, C. & Oswald, A. J. The scientific value of numerical measures of human feelings. PNAS 119(42), e2210412119 (2022).
11. Layard, R., Clark, A. E., Cornaglia, F., Powdthavee, N. & Vernoit, J. What predicts a successful life? A life-course model of wellbeing. Econ. J. 124(580), F720–F738 (2014).
12. Lucas, R. E. Long-term disability is associated with lasting changes in subjective well-being: Evidence from two nationally representative longitudinal studies. J. Personality Social Psychol. 92, 717–730 (2007).
13. Oswald, A. J. & Powdthavee, N. Does happiness adapt? A longitudinal study of disability with implications for economists and judges. J. Public Econ. 92(5), 1061–1077 (2008).
14. Lucas, R. E., Clark, A. E., Georgellis, Y. & Diener, E. Unemployment alters the set point for life satisfaction. Psychol. Sci.
15. 15(1), 8–13 (2004). 15. Kassenboehmer, S. C. & Haisken-DeNew, J. P. You’re fired! the causal negative effect of entry unemployment on life satisfaction. Econ. J. 119(536), 448–462 (2009).